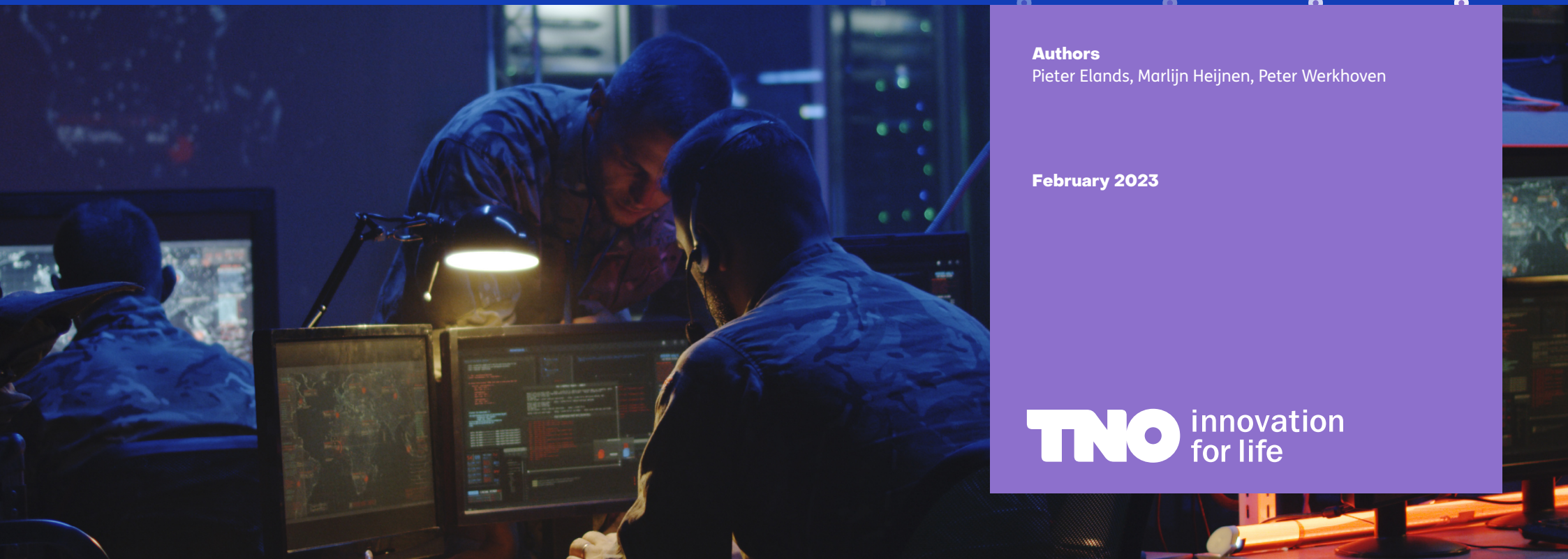
Position paper

Operationalization of meaningful human control for military AI

A way forward



Authors
Pieter Elands, Marlijn Heijnen, Peter Werkhoven

February 2023

TNO innovation for life

Position paper for the REAIM 2023 summit

Introduction

The rapid rise of Artificial Intelligence (AI) has fuelled the development and use of intelligent and autonomous systems in the military domain. As the capabilities of AI increase, the opportunities to do parts of human work and augment human cognition will increase too.

Potential benefits are the advancement of the efficiency, effectiveness and safety of the operations of armed forces. However, serious downsides of AI-deployment can appear that prove to be hard to predict or anticipate. Keeping the human meaningfully in control of this wicked problem of balancing the up- and downsides of AI-development and deployment entails three challenges.

First, the responsibility gap must be addressed at the levels of governance, design, development and operation (i.e., the moral culpability, moral and public accountability, and active responsibility).

Second, AI's objectives and behaviours must be well aligned with the values of the stakeholders at all of these levels.

Third, the dynamic situation ('situated-ness') influences the AI and human performance, and relevance of the values at stake.

To date, there is no international consensus on the definition of meaningful human control (MHC) (AIV/CAVV, 2021)[1] and how to address the responsibility gap, value alignment and situatedness in the development and deployment of AI.

There is agreement on *guiding principles*, such as formulated by the UN Group of Governmental Experts[2] and NATO[3], and there are *analytical frameworks* to identify the problems. However, a comprehensive prescriptive approach is lacking for building and implementing AI-technology in such a way that it is under meaningful human control, during its complete lifecycle.

In this short position paper we present such an approach, *operationalizing* MHC with a continuous 'creation-feedback loop' of developing, controlling, evaluating and adjusting human-AI systems at the level of governance, design, development and operation, the so-called (multi-level) *Socio-Technological Feedback Loop* (STFL). The STFL can be widely used for responsibly employing AI-systems, but will be illustrated here for high-risk military applications, in which high-risk refers to the risk of unintentional harm.

The outcome of the STFL is situationally dependent, that is, it is affected by the specific AI-system deployed and the specific context of the governance, design, configuration and operation. For the main part, the examples of this paper centre on the operation level to illustrate the proposed methodology, while possible governance, design and configuration decision processes, and the corresponding evaluation, feedback and adjustment processes are less worked out. For many situations the mission may be best achieved with 'human-in-the-loop' solutions. But the STFL does not exclude 'human-before-the-loop' solutions for situations for which MHC via 'human-in-the-loop' is not realistic or impossible (e.g. defence against saturation attacks). In that case 'value alignment' and 'responsibility allocation' can be reliably achieved 'before-the-loop' using machine interpretable moral models. The use of moral models for the AI-system can also serve as a decision support tool for 'humans-in-the-loop'.

It is important to note that the STFL yields promising outcomes, but is still heavily under research.

1 AIV/CAVV advice 2021 and cabinet response 2022: https://www.adviesraadinternationalevraagstukken.nl/documenten/publicaties/2021/12/03/autonome-wapensystemen
2 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
3 NATO Principles of Responsible Use: https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html

The Human-AI system

The system to be developed is by definition a collaborative human-AI system with 'system objectives', the objectives to be achieved in its mission. To act effectively in such a 'human-AI' system the AI-system must first have an objective function consistent with the human-AI system objectives. Second, it needs a certain knowledge of the world it operates in (a world model) to be able to reason about the effects of its actions and (socially) interact with humans. Depending on the context and risk of harm of operation, the objective function and world model can include legal and ethical aspects and principles. It is important however, to realize that a world model and a moral model are by definition approximations of reality and that their functionality is situationally dependent. The functioning of humans, AI-systems and the human-AI system as a whole during operation is often described using the OODA (Observe-Orient-Decide-Act)[4] loop.

Socio-Technological Feedback Loop

The Socio-Technological Feedback Loop (STFL) is a methodology to 1) identify the relevant ethical, legal and societal aspects the behaviour of the human-AI system should adhere to, and 2) ensure that the human-AI system operates according to those aspects. There is no single solution that achieves both in all possible applications of AI-technologies. Instead of aiming for such a solution, the STFL is a methodology; a set of methods to identify and operationalize the relevant ethical (legal and societal) aspects (given the mission goals) to establish meaningful human control of a specific AI-system in a specific context.

Other known approaches such as *Value Sensitive Design*[5], ELSA/I, and *Responsible Research and Innovation*[6] share important aspects with the STFL such as stakeholder involvement and multidisciplinary design.

The STFL methodology differs from these, as it incorporates such approaches and places them in a continuous process of improvement. As such, it connects these approaches with each other.

The STFL as illustrated in Figure 1, is a continuous process and forms nested loops with different timescales: the governance (or development) loop, the design loop, the configuration loop, and the operation loop. Each larger feedback loop governs its smaller counter-parts. For example, the governance loop dictates the design process (i.e., guiding principles signifying what should be incorporated into the design), whereas the design process dictates what can be configured (i.e., which behavioural constraints are available) and how the human-AI system should operate (e.g., as supervisory control). All loops include verification and validation. Solutions to establish meaningful human control must be found within these loops.

At the start of a new mission, it may be necessary to go through one or more loops again, to obtain the best possible (and approved)[7] moral behaviour of the human-AI system for that specific mission. The STFL addresses the whole lifecycle of that system; the resulting system behaviour is to be evaluated, verified, validated and, when needed, refined regularly. Changing ethics in society also require regular execution of the governance loop, and possibly its inner loops. Furthermore, new or unforeseen contexts of operation may require adjustments in the design, configuration and operation of the human-AI system.

Although these loops have different stakeholders, they are highly intertwined, and relevant stakeholders must be involved throughout the process.

4 John R. Boyd, Destruction and Creation, US Army Command and General Staff College, 3 September 1976.
5 Friedman, B. (1996), Value-sensitive design, Interactions, 3(6), 16-23.
6 Schomberg, R. von (2011), 'Introduction', In: R. Von Schomberg (ed.), Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields. European Union Publications Office p. 7-15.
7 Current practice foresees in a combination of setting Rules-of-Engagement (by the legislator) and review of compliance with law by legal advisors.

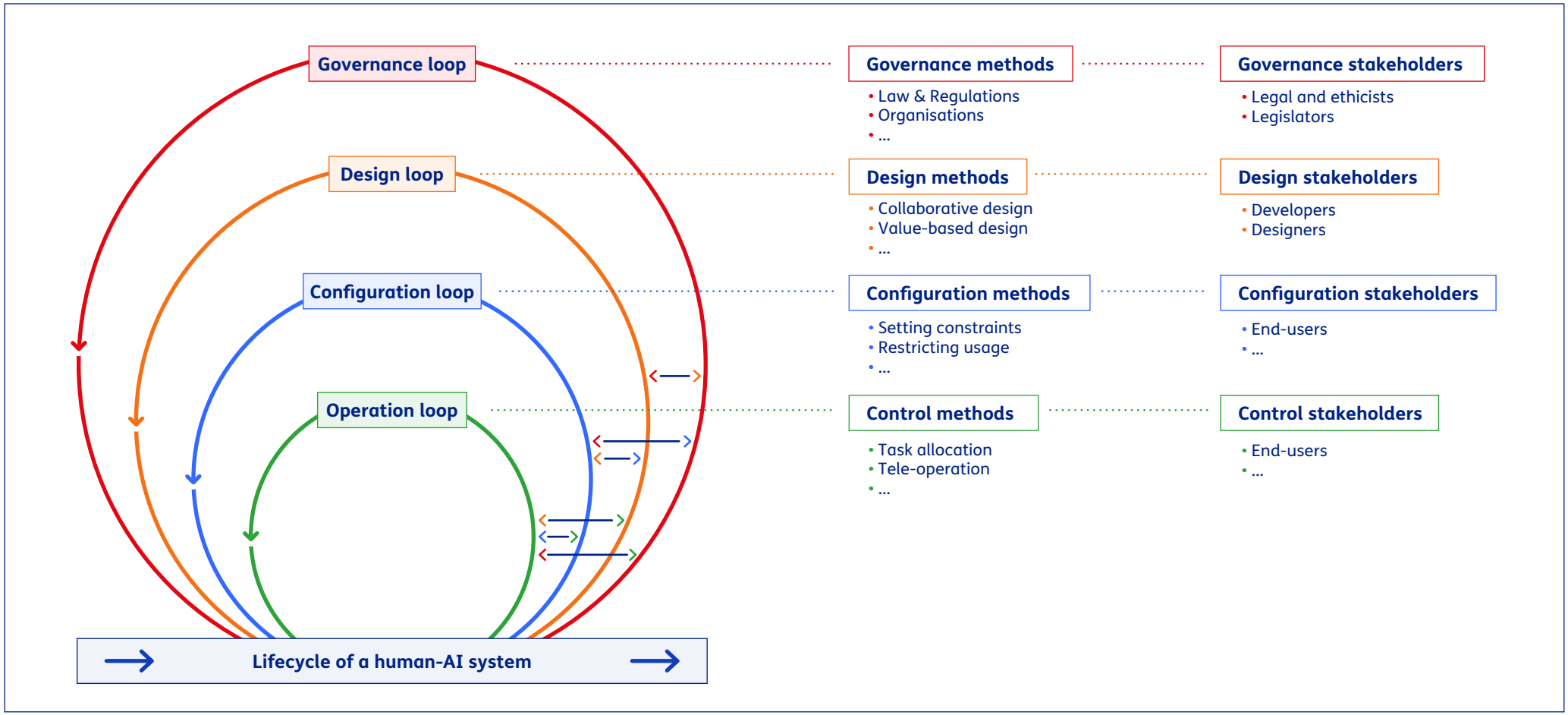


Figure 1: An illustration of Socio-Technological Feedback Loop. The governance loop is a continuous process of arriving at relevant principles, guidelines, policy, laws and similar dictating aspects (e.g., international debate on autonomous weapons). The design loop (e.g., control paradigm of the 'human-AI' system and AI-system based on operational, technical and ethical/legal considerations). In the configuration (or development) loop, the human-AI system is implemented according to the design specifications (allowing to reiterate the design process if requirements give too much room for interpretation). In the operation loop, the human-AI system is applied in the operational context it was designed and developed for (feedback from the operation loop can start new configuration, design and governance loops).

Whereas Figure 1 illustrates the (nested) loops of the STFL, Figure 2 illustrates its (iterative) phasing when put into practice, specifically for the configuration (development) loop. If specified by the previous design loop (not illustrated in Figure 2), it includes the selection of a harm model. The figure illustrates how the STFL methodology can structure concrete methods into a responsible and practical process, including validation. The iterative improvement loop is part of the entire lifecycle.

The STFL starts with the formation of a stakeholder team, with all relevant stakeholder categories at the table and with all stakeholders having a mandate to decide. Examples of stakeholders involved include legal experts, legislators, ethicists, military users, system engineers, AI developers, and NGO's. Some formal entity (likely the government) has to decide which stakeholders must be at the table. The team of stakeholders is given the case of a specific human-AI based system to be used in a specific context. The first task for the team is to assess all potential risks of unintentional harm, caused by the (human-AI) system, operating in that specific context.

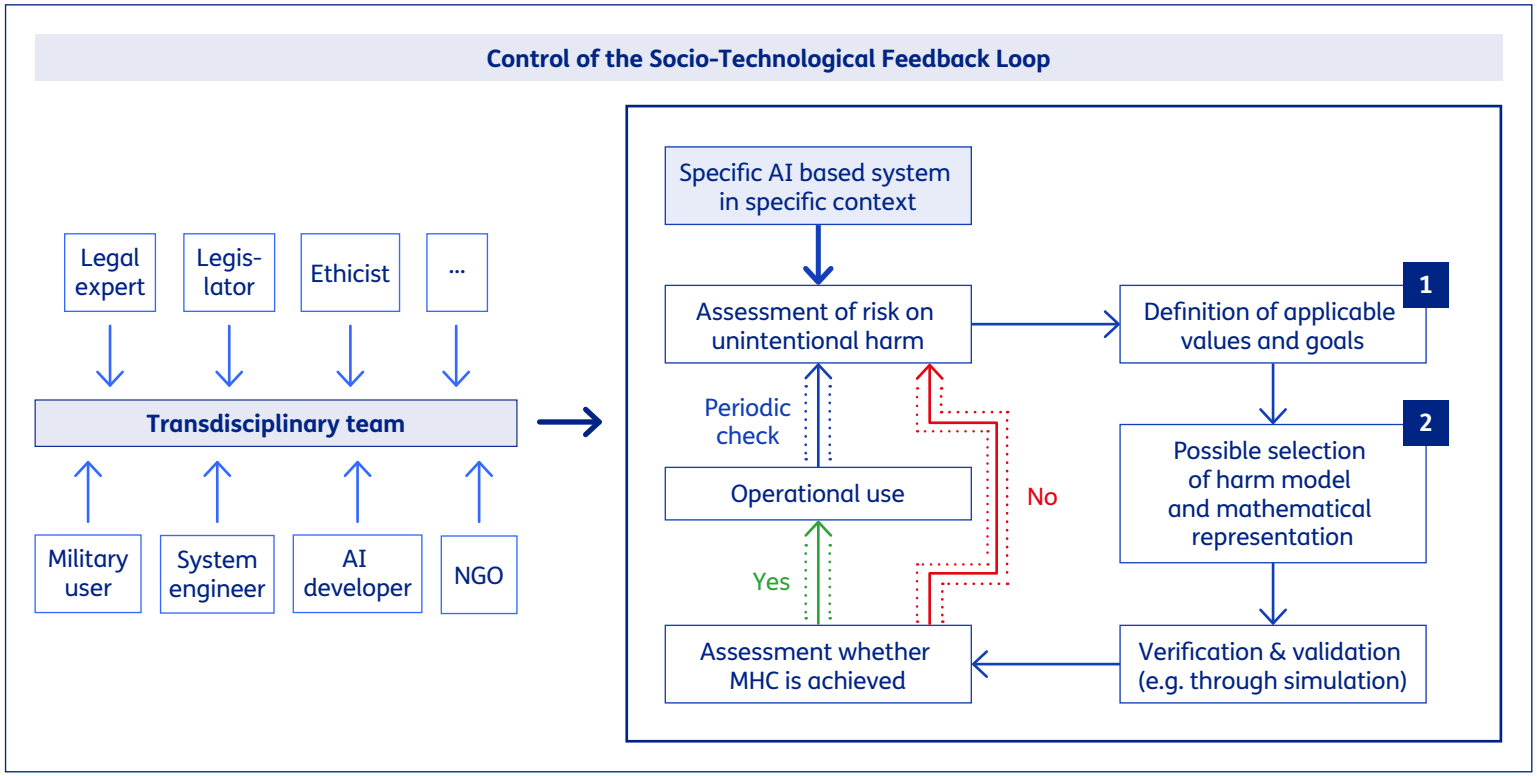


Figure 2: Phasing of the Socio-Technological Feedback Loop.

Next, the team has to define the mission goals and applicable values (step #1 in Figure 2). The team also has to ensure that all these elements are included in the systems' world model. It is as yet unusual for present military operations, to specify these values, as it requires to makes these values explicit in relation to achieving the mission goals, to satisfy proportionality and subsidiarity requirements. Currently only specific constraints, such as Rules of Engagement (ROE) are specified. We see this as an explicit responsibility of the legislator.

The next step for the team, indicated as step #2 in Figure 2, is the decision if and what type of 'harm model' or 'moral model' is used. A harm model (or 'method') is an (approximate) representation of the set of values and goals and may serve as an instruction to the (human-machine) system.

For relatively predictable systems in bound contexts, relatively 'straightforward' methods, such as a set of rules, may suffice; the Goalkeeper Close-In Weapon System in use with the Royal Netherlands Navy may be considered as an example of this. Other situations require a more complex method, such as methods aimed at optimizing human-AI interaction, such as value-based design and usability engineering, especially suitable for complex situations, which do not deviate too much from the situation which the AI was designed for. A trade-off for high-risk situations (complex, time-critical) – between the risk of incomplete ethical value-alignment and the risk of harm caused by *not* deploying AI-systems – may result in 'human-before-the-loop' deployment of the AI-system. In this case, ethical values must be maximally explicitized through a (mathematical) 'harm model'.

A substantial amount of research is still required to successfully achieve operationalization in case of complex systems in complex situations, as the level of risk and the complexity of the context determine the complexity of the harm model as an outcome of the STFL.

Of crucial importance is also the phase of verification and validation, see Figure 2, to assess whether or not the selected approach is able to adequately deal with a large number of situations, in terms of keeping unintentional harm below an acceptable level, as determined by the shared objectives and values of the human-AI system. In this phase, it is also assessed whether the system correctly interprets the world based on its observations, which can be challenging, and whether the AI-system provides the best possible solutions. If this is the case, the system may be declared fit for initial deployment by the team for the specified mission, keeping in mind that to serve the entire lifecycle well, the STFL has to be continuously iterated.

Closing remarks

- There is not yet an institutionalized body to assess if and when a stakeholder team is an adequate representation of all relevant stakeholders.
- The advantage of the STFL is that goals and values are made explicit before the actual mission, not delegating the full responsibility for this to the executing team.

Example in an operational setting

The example below illustrates the functioning of a human-AI system. The application context dictated a human-before-the-loop control paradigm as was determined in the design process. The AI system itself has an advanced model of moral, ethical and legal aspects integrated into its objective and an equally advanced supporting model of the world. This was decided upon through validated methods in the design phase and implemented and tested in the configuration (development) phase. This example depicts how the STFL methodology could achieve responsible application of military AI in the future as research towards it continues.

Scenario: *A military compound operated by UN military in a country with civil war, suddenly notices a flow of inhabitants from the neighbouring village moving towards the compound, still at some distance. They are seeking shelter from insurgents, harassing the village. Meanwhile, the anti-drone radar of the compound signals dozens of small drones, armed with grenades, heading directly for the compound. The commander gives orders to start operating the counter-drone system. She has to take care not to attack the drones flying above the refugees; they could be injured or killed by exploding grenades. The commander launches a drone to have a real-time picture of the situation; this way she is able to only counter the drones which do not form a risk to the refugees.*

Here the prime risks are the unintentional injuring or killing of refugees, the risk of people in the compound getting hit and the risk of damage to the infrastructure. The latter risks are considered as minor, since the compound has armoured buildings where people can take shelter. The challenge for the commander is to maximize the countering of drones while avoiding collateral damage. Despite the situation sometimes being hectic, the outcome is satisfactory. The STFL methodology requires these values and goals to be made explicit in the design loop.

Suddenly the image from the drone disappears, the commander cannot locate the positions of the refugees any more. The commander has three options: (1) continue to counter the drones and risk casualties among the refugees, (2) cease fire and allow the drones to attack the compound, and (3) switch the counter-drone system, which still can receive information from the drone, to fully automatic. The counter-drone system has been designed (governance loop and design loop of STFL, see Figure 2) to not attack drones flying above people, but it is less accurate than a skilled human operator.

Option 1 is unattractive, the risk of civilians getting killed must be avoided. The risk of casualties and damage to the compound will increase in option 2, while the risk of casualties among the refugees increases if option 3 is chosen. However, soon the first refugees will arrive at the compound, which implies that option 2 will cause an increasing risk for the refugees, who cannot find immediate shelter. In the STFL approach, the weighing of options is a continuous process, and may take place in all loops.

The commander decides to put the counter-drone system on fully automatic (configuration loop and operation loop in STFL, see Figure 2) and gives orders to help the refugees find shelter. For a while, this works. Refugees which have entered the compound safely, report that a large number of heavy vehicles, including rocket artillery vehicles, commanded by the insurgents, are gathering in the village. They could pose a serious threat to the compound and all its inhabitants. Should the commander direct the drone to observe the village and stop countering the drones?

Observations: A number of observations can be made. First, all possible risks of unintentional harm have to be assessed and have to be given a value. In case new risks come up, the objective function and the world model have to be extended, if necessary and if possible, to include the new objects, phenomena, aspects, etc., involved. If the objective function and the world model do not cover the present situation, the human operator must take over control and decide to continue or not. Second, the minimization of total unintentional harm is dependent on the specific situation; different situations may result in different outcomes. Every change in the situation requires a new optimization of results and minimization of unintentional harm.

Conclusion

The Socio-Technological Feedback Loop methodology comprises the assessment of a specific human-AI system operating in a specific context through a transdisciplinary and multi-stakeholder approach. Moral, ethical and legal aspects as well as objectives for this human-AI system in high-risk situations are made explicit, comparable, and auditable. It provides a clear attribution of responsibility and accountability; as such it is a way forward to operationalize meaningful human control of military AI based systems. It challenges the stakeholders to make explicit and validate their goals and moral values, for the specific context the system is to operate in.

Authors

Pieter Elands
Program Manager

pieter.elands@tno.nl

Marlijn Heijnen
Scientist

marlijn.heijnen@tno.nl

Peter Werkhoven
Chief Scientist TNO

peter.werkhoven@tno.nl

TNO innovation for life

Context

This publication was submitted by invitation to the organization of REAIM 2023, the first global summit on responsible artificial intelligence in the military domain, hosted by the Dutch government.

tno.nl/responsibledecisionAI