Governing ethical and effective behaviour of intelligent systems

A novel framework for meaningful human control in a military context

Unmanned systems are gaining a permanent position in the military domain; they are deployed where people are physically inadequate or at risk. Increasingly often these systems can also perform tasks independently, made possible by the strong advance of artificial intelligence. On cognitive tasks such as dealing with large amounts of data, understanding complex problems and rapid decision-making, they increasingly surpass people. However, this trend towards more autonomy raises the question: how do we as human beings maintain control over autonomous systems, who exerts that control and how do we justify it? In this article TNO presents a new framework for meaningful human control of autonomous (intelligent) systems.

Although technology development in AI is led by the civil domain, military application of AI increases rapidly

PHOTO U.S. AIR FORCE, BARRY LOO

*Ir. P.J.M. Elands, Ir. A.G. Huizing, dr. ir. L.J.H.M. Kester, dr. M.M.M. Peeters and dr. S. Oggero**

The introduction of intelligent systems[1] for military operations raises ethical issues about maintaining meaningful human control. Advocates of a ban on autonomous weapon systems sketch visions of a future where robots define their own targets and decide who lives and who dies. In our opinion such decisions must always be made under human control. In addition, in the current practice of military operations, effective and meaningful human control is sometimes difficult to achieve. This article addresses the question of meaningful human control: 'how can we exploit the benefits of intelligent systems in military operations, while ensuring ethical behaviour, and effective, safe and responsible operations?'

This article starts with an overview of the developments in autonomy and artificial intelligence and their significance for military applications, including the international debate on autonomous weapon systems. Next, the article discusses a number of challenges in the current practice of military operations, which may be addressed by the use of autonomy and by a proper framework for meaningful human control. The main body of the article consists of a description of a framework for meaningful human control of autonomous weapon systems and addresses the ethical guidelines, the use of a goal function, roles and responsibilities, trust and uncertainty, accountability, and the advantages of the framework. A short research agenda and conclusions finalize this article.

The rise of intelligent systems

Over the last few decades technological advancements have rapidly changed the way we work and live. Developments in Artificial Intelligence

Governing Ethical and Effective Behaviour of Intelligent Systems

Lieutenant colonel Royal Netherlands Air Force W. Ligtenberg, MSc EMSD

'All intelligent organisms and organizations undergo a continuous cycle of interaction with their environment,' said John Boyd, an important systems thinker of the past. In fact, Boyd argues that the military also needs to think about the challenges of today and tomorrow using a systems approach, and rightfully so. Within the military we are confronted with new dynamic and complex surroundings. This increase in complexity can only be fully understood using a more holistic approach. New domains like cyber force us to be knowledgeable on more than the physical battleground alone. But is it not true that most military organizations are already challenged to effectively conduct decision-making for current combined or joint operations? What does it mean to orchestrate effective decision-making for military operations within a true multi-domain environment?

The amount of data being generated by these domains can easily be overwhelming. Furthermore, effective decision-making requires ever shrinking timespans. There is also reason for concern that the fastest decision-making process, linked with the ability to directly create (non-) kinetic effects, will generate an important advantage during conflict. Although we have some supporting analytical systems in place, it seems that our best attempt to generate a holistic approach by fusing data is not yet quite good enough. Often commanders experience fusion of data to support their decision-making in a complex environment as a cumbersome process. It is inevitable that we as a military think about new ways to structure and support decision-making. We need to find solutions that will empower us to achieve more effectiveness in our operations, adhering to shrinking timespans for decision-making, but without ignoring the checks and balances that need to be in place.

In this article the authors explain how this effectiveness can be reached if we are willing to understand that augmented decision-making – using intelligent systems – can also be a very reliable solution. The importance of this discussion cannot be underestimated. Together, scientists and military experts first need to conceptualize relevant frameworks for an intelligent system. These frameworks can be used to have more informed and broader – multidisciplinary – discussions, and at the same time start building responsible solutions.

* Pieter Elands, Albert Huizing, Leon Kester, Serena Oggero and Marieke Peeters work at TNO Defence, Safety, and Security in the Netherlands. Their research encompasses the possibilities of artificial intelligence and its application in autonomous systems in various programmes for the Dutch Ministry of Defence.

1 As explained later, we prefer the term 'intelligent systems' instead of 'autonomous systems'.

A Global Hawk returns to base: military systems such as UAVs are now being perceived as 'killer robots'

(AI) and computational processing power have given rise to increasingly intelligent systems, i.e. entities capable of engaging in dynamic and goal-directed interaction with their environment, using some form of AI.[2] Intelligent systems are often described in terms of their behaviour and capabilities: intelligent systems can sense their environment, reason about their observations and goals in order to make decisions, and act upon their environment.[3] Intelligent systems outperform humans in handling large amounts of heterogeneous data, dealing with complex problems, and rapid decision-making.[4] Continuing improvements in AI, e.g., machine learning and problem solving, and computational processing power are expected to further enhance these advantages for decades to come. The result of these developments is that nowadays an increasing number of tasks is carried out by intelligent systems. A more recent development is that, for certain tasks, intelligent systems are capable of operating at high performance levels for extended periods of time without the constant need of human support or intervention, leading to the use of the term 'autonomous systems'.[5]

Although technology development in AI is led by the civil domain, military application of AI

2 P.D. Scharre and M.C. Horowitz, *Artificial Intelligence: What Every Policymaker Needs to Know* (Washington, D.C., Center for a New American Security, 2018).

3 M. Wooldridge, An Introduction to Multi Agent Systems (New York, Wiley, 2009).

4 A.P. Williams and P.D. Scharre, *Autonomous Systems – Issues for Defence Policymakers* (Norfolk, Virginia, NATO HQ SACT, 2015).

5 M. Vagia, A. Transeth and S. Fjerdingen, 'A Literature Review on the Levels of Automation During the Years. What are the Different Taxonomies that have been Proposed?', in: *Applied Ergonomics* 53 (2016) 190-202; J.M. Beer, A.D. Fisk and W.A. Rogers, 'Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction', in: *Journal of Human-Robot Interaction* 3 (2) (2014) 74-99; M. Johnson, J.M. Bradshaw, P.J. Feltovich, R.R. Hoffman, C. Jonker, B. van Riemsdijk and M. Sierhuis, 'Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design', in: *Intelligent Systems* 26 (3) (2011) 81-88; R. Murhpy and J. Shields, *The Role of Autonomy in DoD Systems* (No. 20301-3140) (Washington, D.C., Defense Science Board, 2012).

PHOTO U. S. AIR FORCE, CHAD BELLAY

increases rapidly. Since a couple of years intelligent systems have been considered to provide a decisive advantage on the battlefield;[6] for that reason many countries are actively developing military systems with such capabilities.

In our view, the term 'autonomy' in relation to intelligent systems is confusing, as it is often wrongly interpreted that such systems determine their own ethical goals, autonomy in the sense of 'making decisions without being controlled by anyone else'. Yet these systems are in fact bounded by the tasks and/or goals assigned within a legal framework by their manufacturer, owner, and/or operator. Military systems, such as UAVs or 'drones', are now being perceived as 'killer robots', sketching visions of a future where robots define their own targets and decide who lives and who dies.[7] We agree that such choices should always be made by humans.

Concerns about the proliferation of intelligent systems in the military, as well as other high-risk domains, has led researchers and companies in the fields of robotics, ethics, philosophy, and artificial intelligence to sign open letters in which they plea for firstly a prioritization of research on robust and beneficial artificial intelligence, secondly a ban on autonomous weapons systems, and thirdly refraining from any activities contributing to the realization of autonomous weapons systems.[8]

The UN Convention on Certain Conventional Weapons (CCW) is the official forum where the debate on Autonomous Weapon Systems takes place. Central to this debate is the term 'meaningful human control'.[9] A big problem in this debate, however, is the lack of consensus about the definition of what is 'autonomy', making it quite difficult what to ban or not to ban.[10] And although there are various definitions of the term 'meaningful human control', there seems to be agreement on the necessity to ensure that intelligent systems will not go beyond boundaries set by humans.[11]

6 R.A. David and P. Nielsen, *Defense Science Board Summer Study on Autonomy* (Washington, D.C., 2016).

7 *Losing Humanity: the Case Against Killer Robots* (New York, Human Rights Watch, 2012); 'Killer Robots and the Concept of Meaningful Human Control', *Memorandum to the Convention on Conventional Weapons (CCW) Delegates* (New York, Human Rights Watch, 2016); *The Problem* (Washington, D.C., Campaign to Stop Killer Robots, 2018).

8 *Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter* (Boston, Future of Life Institute, 2014); *Autonomous Weapons: an Open Letter from AI & Robotics Researchers* (Boston, Future of Life Institute 2015); *Lethal Autonomous Weapons Pledge* (Boston, Future of Life Institute, 2018).

9 B.L. Docherty, Human Rights Watch, Harvard Law School, and International Human Rights Clinic, *Making the Case: The Dangers of Killer Robots and the Need for a Pre-Emptive Ban* (New York, 2016); Article 36, 'Key Areas for Debate on Autonomous Weapon Systems' (2014); Article 36, 'Killing by Machine – Key Issues for Understanding Meaningful Human Control' (2015); Human Rights Watch, *Killer Robots*.

10 M. Ekelhof, 'Autonome Wapens. Een verkenning van het concept Meaningful Human Control', in: *Militaire Spectator* 184 (2015) (5) 232-245.

11 In the light of this debate, various reports have been published by research groups and institutes, e.g., P.D. Scharre, *Artificial Intelligence* (2018), G. Schaub and J.W. Kristoffersen, *In, on, or out of the loop?* (2017), Centre for Military Studies – University of Copenhagen; V. Boulanin and M. Verbruggen, *Mapping the Development of Autonomy in Weapon Systems* (2017), Stockholm International Peace Research Institute; and R.A. David et al, *Summer Study on Autonomy* (2016). In addition, new research initiatives have been raised, such as *AiTech – Meaningful Human Control of Autonomous Intelligent Technology* (2018), Delft University of Technology; Future of Life Institute (2014), and CLAIRE (Confederation of Laboratories for Artificial Intelligence in Europe) (2018).

As part of the debate, the concepts 'human-in-the-loop', 'human-on-the-loop', and 'human-out-of-the-loop' are often used to clarify what is and what is not desirable to establish meaningful human control.[12] In our view these terms confuse the debate even more, as they seem to imply that: (1) there is exactly one human who (2) either is or is not part of (3) an unspecified control loop. Instead, we argue that – depending on the situation at hand – the system may allow for various types of human control by various people in various roles. Furthermore, we emphasize that meaningful human control is a combination of measures complementing one another to ethically and effectively govern the behaviour of intelligent systems.

Because of the problematic notion and strongly polarised opinions on 'autonomous systems', this paper prefers to use the term intelligent systems: systems that are capable of independently performing specific tasks within specific contexts for specific time periods, yet also accept specific types of control from particular people in predefined roles.

Developments in AI will continue at a rapid pace and AI will continue to find applications in the military domain. Military systems will become better and smarter. It is our belief, however, that mankind should always be in control of its technological inventions and hence should not allow intelligent systems to set their own goals. For that reason this article presents a novel framework for meaningful human control.

Observations and issues in current military practice

Looking at the current practice of military operations, there are some observations to make. First of all, ethical and legal guidelines provided, e.g., Rules of Engagement (ROEs), from the political/strategic level, often need further interpretation and consideration at the operational and tactical level. Sometimes this has to be done under significant time pressure, which is something that humans are not necessarily good at.[13] It is therefore not strange that military personnel in specific situations can feel uncertain or uncomfortable making ethical decisions. We believe these considerations and decisions must lie with people rightfully entitled to do so, i.e. the legislative power.

A second observation is that in current military operations the connection between actions, effects, and mission goals has often neither been explicated nor quantified.[14] The commander translates the mission goals and desired military end state to effects to be achieved and actions to be taken in a qualitative manner. More often than not, this results in performance measures that are actually more related to the effort made, such as number of sorties or number of munitions dropped, instead of the effects obtained. As a result, the mission goals may get out of sight, especially at times when time pressure or other stress factors affect the decision making process. In contrast, when using an intelligent system, formalized quantitative relations between mission goal, effects and actions are mandatory; the system has to know how specific actions and their effects relate to the mission goal.

A third observation is that one could argue whether the current way of conducting military operations fulfils the requirements of 'meaningful human control'.[15] Because of the complexity of military operations, the process of control is distributed into many different sub-processes. It is important to note that part of these sub-processes have already been automated, such as collateral damage estimation, to better accommodate their complexity.[16] In cases like these,

12 Schaub et al, *In, on, or out of the loop*.

13 D. Kahneman and P. Egan, *Thinking Fast and Slow* (New York, Farrar, Straus and Giroux, 2011).

14 J. van Deventer, 'Meten is weten, maar wat meten we eigenlijk?', in: *Think Airpower. Newsletter Air and Space Warfare Center*, no. 9 (2015).

15 M. Ekelhof, 'Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control through the Lens of Military Targeting', in: *Naval War College Review* 71 (3), article 6, 2018.

16 Boulanin and Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*; R. Danzig, *Technology Roulette – Managing Loss of Control as Many Militaries Pursue Technological Superiority* (Washington, D.C., Center for a New American Security, 2018).

The Sea Hunter recently completed an autonomous sail from San Diego to Hawaii and back: because of the complexity of military operations, the process of control is distributed into many different sub-processes

PHOTO U.S. NAVY

the designers of the underlying algorithms have determined what information is important to take into account, what information can be aggregated into a single representation, what information is considered as 'noise', and how uncertainty should be measured and taken into account. Such algorithmic decisions at the lower level may unwittingly affect the outcomes of human decision makers. The success of a mission, then, depends on a collection of individuals, both human and artificial, to perform their part of the task. Yet due to this distributed control process it is hardly possible for a single person to maintain a proper understanding of, in the first place, each individual's unique contribution to the ultimate outcome, as well as, secondly, the dynamics and interdependence between each of the individuals in fulfilling the mission goal. In other words, this complex process confuses the attribution of responsibility for the ultimate outcomes. This is often referred to as the 'many hands problem'.[17] In the light of these findings, it may be concluded that even in the current situation effective and meaningful human control is already quite a challenge.

Framework for meaningful human control

The desire for meaningful human control when introducing intelligent systems combined with the observations on human control in current military practice leads us to the main question on what, in our view, meaningful human control is truly about: '*how can we exploit the benefits of intelligent systems in military operations, while ensuring ethical behaviour, and effective, safe and responsible operations?*'

Ethics – A Normative Approach

A popular ethical framework for meaningful human control is a normative approach, in which laws and ethical norms are embedded in

17 D.F. Thompson, 'Moral Responsibility of Public Officials: the Problem of the Many Hands', in: *American Political Science Review* 74 (1980) (4) 905-916; D.F. Thompson, 'Responsibility for Failures of Government: the Problem of Many Hands', in: *The American Review of Public Administration* 44 (2014) (3) 259-273; I. van de Poel, L. Royakkers and S.J. Zwart, *Moral Responsibility and the Problem of Many Hands* (New York and London, Routledge Taylor & Francis, 2015).

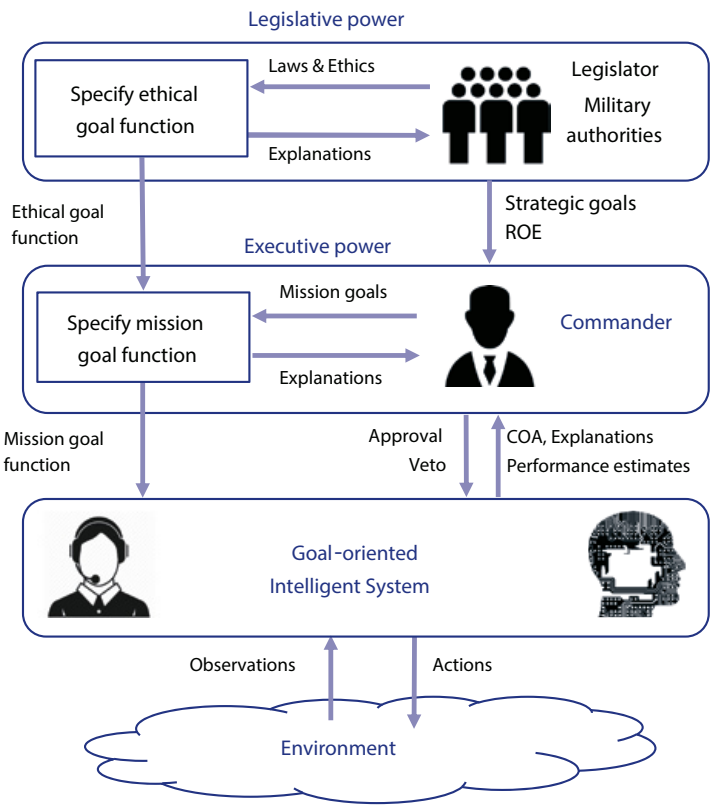

Figure 1 Meaningful human control of an intelligent system in a military context through an ethical goal function and a mission goal function

an intelligent system.[18] In a military context the normative approach could be implemented by embedding the ROEs for a military operation in an intelligent system.[19] The benefits of this approach are that the ROEs are transparent and easily interpretable for humans and that it is generally clear when ROEs have been violated. However, a normative approach requires that there are rules for every situation that can be encountered, which is not easily feasible in practice. Moreover, there are situations in which conflicting rules apply, which can only be solved by a trade-off between the utility of actions.

Ethics – A Utility-Based Approach Using Goal Functions

As an alternative to the normative approach this article proposes a utility-based ethical approach, using a framework for meaningful human control that obliges intelligent systems to take the same decisions as qualified people, with sufficient decision time and access to relevant information, would make in the same situation. This objective is achieved by providing the intelligent system with so-called goal functions that represent the ethical values and military goals and constraints of the mission as defined by the legislator, military authorities and military commander;[20] see Figure 1. A goal function is a mathematical function in which several different goals are combined. As an example may serve an autonomous vehicle which has to transport a passenger from A to B. In this case several different goals are involved, such as travel time, passenger comfort, environmental impact, and road safety. Different weight factors may apply to these different goals. The autonomous vehicle will use the goal function to decide upon the route to take and upon its driving behaviour to best take into account both the wishes of the passenger (comfort, time of arrival) and of society (environment, safety), closely resembling the human decision making process. Depending on the weights and on the present state of the world (amount of traffic), different outcomes are possible. This approach contrasts with the current military practice of predefined desired end states.

Roles and Responsibilities

This goal-oriented control concept requires a machine-interpretable, mathematical goal function, which combines ethical and legal goals with specific military mission goals. The ethical and legal goals are specified by the legislative power and include the laws of war. These goals are preferably generally applicable to any type of military operation. They reflect the ethical and legal values of society and are

18 V. Bonnemains, C. Saurel and C. Tessier, 'Embedded Ethics: Some Technical and Ethical Challenges', in: *Ethics and Information Technology* 20 (2018) 41-58; *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2 (New York, IEEE, 2017).

19 A. Cole, P. Drew, R. McLaughlin and D. Mandsager, *San Remo Rules of Engagement Handbook* (San Remo, International Institute of Humanitarian Law, 2009).

20 Also referred to as 'utility functions'.

also referred to as the 'social contract' or 'society-in-the-loop'.[21] They should, due to developing society and technology, continuously be updated through a 'socio-technological feedback loop'.

When a specific mission is sanctioned at the political/strategic level, the government, together with the military authorities, specifies the strategic goals for that mission, including the military goals. In addition, the government and/or the military authorities specify constraints, i.e. the ROEs.[22] The strategic goals and ROEs for this particular mission are passed on to the commander.

The commander uses the strategic goals and ROEs to specify the mission goal function, see Figure 1. This mission goal function must be compliant with the ethical goals specified earlier by the legislator and must adhere to any applicable legal boundaries and ROEs. The mission goal function provides a value to every possible transition of the world, due to outcome of actions. It is important to note that, in contrast to the conventional, task-oriented approach used presently, individual goals are no longer represented as pre-selected desired end states, but as elements of the goal function, each with a specific weight factor, as in the transport example mentioned earlier. A very important goal will therefore weigh heavily.

In case a military mission has to be defined, without an ethical goal function made available, the commander may, under the pressure of time, define the ethical and legal aspects of the mission goal function himself, allowing the legislative power to evaluate his choices.

Trust and Uncertainty

The mission goal function is used to govern the intelligent system, which aims to obtain an outcome which best fits the mission goal function, taking into account the weights of the individual goals. It uses its observations to make an estimate of the current state of the world, i.e. the situation of the world including the system itself, and of the changes in that state resulting from possible actions which could be taken by

21 I. Rahwan, 'Society-in-the-Loop: Programming the Algorithmic Social Contract', in: *Ethics Information Technology* 20 (2018) 5-14; B. de Graaf and M.B.A. van Asselt, 'Veiligheid als sociaal contract', in: *Magazine Nationale Veiligheid en Crisisbeheersing* 11 (2013) (6); N.M. Aliman and L.J.H.M. Kester, 'Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations', in: *Delphi – Interdisciplinary Review of Emerging Technologies*, 1 (2019) 23-29.

22 A. Cole et al, *San Remo Rules*.

Performance estimates provided by the intelligent system will enable the commander to assess how well the mission goals are being achieved

PHOTO U.S. ARMY/AURORA FLIGHT SCIENCES

the system. The actions which lead to the most preferable, i.e. most utile outcome – the best course of action (COA) – will be selected and carried out by the system.

To build trust, the commander can ask the intelligent system for additional explanations about the potential consequences of the mission goal function in hypothetical scenarios, e.g., alternative future developments of the mission.

Furthermore, as the effects of actions are often uncertain, e.g., a shot fired has a certain probability of hitting the target, multiple possibilities are taken into account when computing the effects of an action. To deal with such uncertainties, the potential outcomes of a single COA are aggregated across the sequence of actions. Various aggregation algorithms exist, such as optimizing the worst case, i.e. lowest possible utility, or taking the average 'expected utility'. When computing and selecting the best COA, the algorithm compares the various possible COAs with regard to these aggregated effects. The selected COA is then explained to the commander before it is executed, allowing the commander to approve or veto it. Once the COA is being executed, performance estimates provided by the intelligent system enable the commander to assess how well the mission goals are being achieved. Should the intelligent system perform well below expectations or – e.g., due to malfunctioning components – conduct actions that are not compliant with the mission goal function, the commander or an operator can abort the mission. However, if someone decides to intervene against the plans or actions derived by the machine from the goal function(s), that person will be held accountable for disregarding the boundary conditions set by the legislator and/or military authorities, and may need to justify himself to the judicial authorities, see also Figure 2.

Advantages of the Framework

In the heat of the moment it may be difficult for a commander to balance the effectiveness of the planned COA and its compliance with international law and ethics, even when assisted by legal advice. The executive power, the commander, in an armed conflict can be held accountable for the illegal use of force. The goal-oriented control concept anticipates potential dilemmas that may occur during the mission by clearly and quantifiably specifying the mission goal function, which includes legal and ethical aspects in advance of the mission at the right level of responsibility. As a result, it can be prevented that an inordinate amount of moral pressure is placed on military commanders and the human operators, possibly leading to over-cautious decisions and less effective actions. Moreover, the goal-oriented concept circumvents the increased risk of human error resulting from cognitive bias due to time pressure and the fog of war in many military operations.[23] Hence, goal-oriented control of an intelligent system in a military context alleviates much of the military commander's burden of making effective and moral decisions.

Furthermore, the requirement of an explicit formalisation of the contribution of actions and effects to the outcome of the goal function in a mathematical relation ensures that all actions contribute maximally to the objectives specified by the goal function, potentially leading to more effective actions and desired effects, so that well-defined performance measures can be established.

Another advantage of the goal-oriented approach is the separation of the system's goal function (the 'what') from its problem solving capabilities (the 'how'). In traditional task-oriented systems, the 'what' is often implicitly defined in the problem-solving code of the system, i.e. hard-coded. As a result it may be difficult to update the system's objectives after its deployment, making the system less flexible to accommodate to new missions or to societal developments as time passes. This may cause the objectives to become outdated, without the possibility to easily update them. Moreover, the objectives the system aims to attain are generally designed and implemented by the manufacturer and its programmers, instead of by the legislative power.

23 Kahneman and Egan, *Thinking Fast and Slow*; Ekelhof, 'Lifting the Fog'.

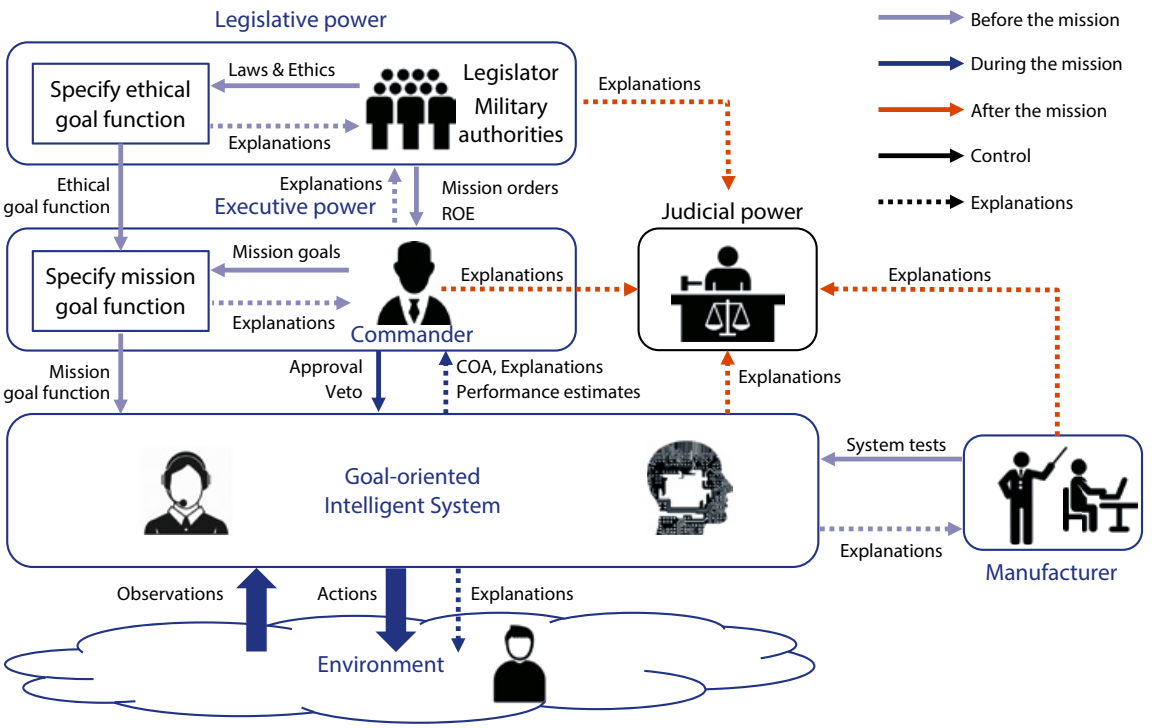

Figure 2 Governance of a goal-oriented intelligent system: the various means of control and explanations before, during, and after of a mission

And lastly, by demanding explicit ethical guidelines defined by the legislative power, these can be inspected and reviewed by the public, e.g., NGOs, NATO, international partners, and/or the judicial power. It might even be possible to use new innovations regarding secure communication, such as blockchain, to prohibit malicious persons from interfering with the ethical goal function provided to the system. To avoid the misconception of malevolent intelligent systems that pursue their own goals, also referred to as 'killer robots', it must be emphasized here that the intelligent system must not be allowed to change its goal function.

Accountability

Figure 2 illustrates the roles of the legislative, executive, and judicial powers in the governance of a goal-oriented intelligent system to clarify the attribution of accountability in case of erroneous or illegal behaviour by the intelligent system during the mission. Figure 2 also shows the various means of control and explanations before, during, and after the mission. It introduces two additional parties to the ones presented in Figure 1, namely the judicial power and the manufacturer. The responsibility of the judicial power is to judge who is accountable for unlawful activities based on the evidence and explanations provided by the executive power (the commander), the manufacturer, and the intelligent system with respect to the pertinent ethical and mission goal functions. To enforce compliance with the ethical and mission goal functions, if the judicial power decides that unlawful activities have taken place, the person or persons responsible for those unlawful activities can be traced back. The manufacturer of the intelligent system is responsible for validating and verifying the proper functioning of the intelligent system by conducting system tests before it is deployed. Such system tests should indicate the system's performance with respect to the ethical goal function, as well as a variety of mission goal functions and scenarios provided by the military authorities.

PHOTO U.S. DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

The new framework provides for a clear distinction between the setting of ethical guidelines by humans and calculating the best course of action by the intelligent system

An interesting option is to combine this goal-oriented approach with a normative approach. This implies that some rules, such as the ROEs, are hard-coded, to set a 'red line'. This may in theory lead to less utile solutions, but it will enhance human trust in intelligent systems and humans will better and more quickly understand their behavior.

Research agenda

The previous section presented a new framework for meaningful human control. It was argued that this framework results in four major improvements. Firstly, the public transparency of the goal function allows for screening by for instance Non-Governmental Organizations (NGOs). Secondly, the framework specifies a clear separation between the goal function and the problem solving capabilities of the intelligent system, also referred to as the orthogonality-based disentanglement.[24] It means a clear distinction between the setting of ethical guidelines by humans and calculating the best course of action by the intelligent system. Thirdly, the framework encompasses a direct link between actions, effects and goals of the mission, allowing for easy determination of the mission effectiveness. Finally, the framework puts the responsibility for defining the ethical values at the proper level.

These claims, however, remain to be validated through scientific research. We therefore propose a research agenda to investigate the value of the proposed framework when applied to real world problems. The first part of this agenda concerns the conception, verification, and validation of ethical and mission goal functions. In the coming year so-called choice model experiments will be initiated.[25] The second part concerns the verification, validation, and optimization of the behaviour of intelligent systems, consisting of humans and machines, in their performance towards the ethical goal function, especially when taking learning capabilities into account. Finally, these research efforts must be complemented by development, building, testing, verification, and validation of intelligent systems, e.g., to assess robustness and

24 N.M. Aliman, L.J.H.M. Kester, P.J. Werkhoven and R. Yampolskiy, 'Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems', Conference on Artificial General Intelligence, 2019; N. Bostrom, 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents', in: *Minds and Machines* 22 (2012) (2) 71-85.

25 C.G. Chorus, B. Pudane, N. Mouter and D. Campbell, 'Taboo Trade-Off Aversion: A Discrete Choice Model and Empirical Analysis', in: *Journal of Choice Modelling* 27 (2018) 37-49; U. Liebe, J. Meyerhoff, M. Kroesen, C.G. Chorus and K. Glenk, 'From Welcome Culture to Welcome Limits? Uncovering Preference Changes over Time for Sheltering Refugees in Germany', in: *PLoS ONE* 13 (2018) (8). BLZ?

scalability. New methods that allow for such activities may need to be developed in the process.

Conclusions

The rise of intelligent systems and their application in the military domain has resulted in an extensive and sometimes heated debate on the desirability of such systems. Various organisations plea for a ban on intelligent systems. This article has observed that misconceptions play a significant role in the debate, and, additionally, that control of 'traditional' military operations has become a distributed and partially automated process, often difficult to oversee.

By introducing a novel framework for ethical decision making, the article hopes to eliminate some of the misconceptions in the current debate on meaningful human control of intelligent military systems and to contribute to effective control of current and future military operations.

By placing an emphasis on specifying a machine-interpretable goal function, humans and intelligent systems are provided with a clear framework, including legal, ethical, and military guidelines. The intelligent system must never be allowed to change this function to set its own goal functions. The mission goal function allows the determination of the value of actions to be carried out in terms of mission effectiveness, which may lead to more effective actions and more desirable effects.

The proposed framework also implies a clear separation of the powers that govern the behaviour of an intelligent system. The legislative power is responsible for the legal and ethical aspects of the goal function, the executive military power is responsible for the military aspects of the goal function, and the engineers are responsible for building intelligent systems that submit to the goal function. The judicial power is responsible for ensuring that all powers honour the ethical goal function as a directive and moral compass. Publication of a

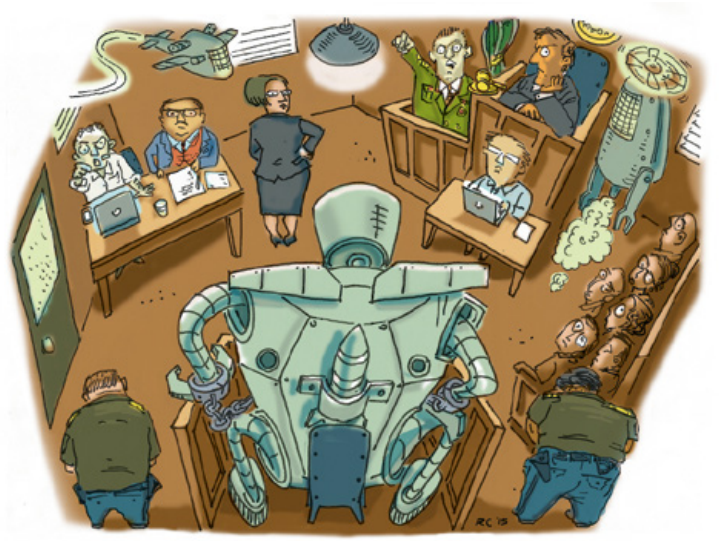

ILLUSTRATION HUMAN RIGHTS WATCH, RUSSELL CHRISTIAN

The judicial power will be responsible for ensuring that all powers honour the ethical goal function as a directive and moral compass

nations' ethical goal function provides transparency to the public and facilitates oversight by NGOs and other stakeholders.

Control of military operations is difficult by nature, due to the complexity of the environment, the fog of war, and the pressure of time. Both human and machine have specific qualities in achieving situational awareness and control. The framework presented in this paper allows for different compositions and levels of cooperation in human-machine teams to achieve optimal results, both in terms of effective control and achievement of the mission goals.

The aim of the proposed framework is to be able to exploit the benefits of intelligent systems for military operations, while ensuring ethical behaviour and effective, safe and responsible operations. It must be realized, however, that the proposed framework requires further substantiation and validation through continuous efforts in line with the proposed research agenda. This cannot be established by scientists and engineers alone, so addressing the mentioned challenges must ideally be done together with subject matter experts at the military and governmental level in a concerted effort. ■