HUMAN–MACHINE INTERFACES IN AUTONOMOUS WEAPON SYSTEMS

Considerations for Human Control

IOANA PUSCAS

ACKNOWLEDGEMENTS

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This study was produced by the Security and Technology Programme, which is funded by the Governments of Germany, the Netherlands, and Switzerland, and by Microsoft. The author wishes to thank the following individuals for their invaluable advice and assistance on this paper: Dr. Giacomo Persi Paoli (UNIDIR), Sarah Grand-Clement (UNIDIR); the following **experts interviewed for this project**: Dr. Mennatallah El-Assady, Dr. Mica Endsley, Parrish Hanna, Dr. Ming Hou, Dr. Matthew Johnson, and three other anonymous experts; and the following **external reviewers** of this study: Dr. Jurriaan van Diggelen, Dr. Marcel Baltzer, Dr. Elisabeth Hoffberger-Pippan, and Prof. Duncan Brumby.

ABOUT UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

NOTE

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in this publication are the sole responsibility of the author. They do not necessarily reflect the views or opinions of the United Nations, UNIDIR, its other staff members or its sponsors.

ABOUT THE AUTHOR

Ioana Puscas is Researcher on AI in the Security and Technology Programme at UNIDIR.

www.unidir.org | © UNIDIR 2022

Photos: © Shutterstock: Cover: Color4260, p2: Gorodenkoff, p8: Hunter Bliss Images, p12: DR MANAGER, p19: ConceptCafe, p21: Jackie Niam, p25: ikatwm.

Table of contents

Abbreviations & Acronyms

AI	artificial intelligence
AWS	autonomous weapon systems
CCW	Convention on Certain Conventional Weapons
GGE	Group of Governmental Experts
HMI	human–machine interface
IHL	international humanitarian law
ISO	International Organization for Standardization
LAWS	lethal autonomous weapons systems
ML	machine learning
UNIDIR	United Nations Institute for Disarmament Research
XAI	Explainable AI

Executive summary

Human control over autonomous weapon systems (AWS) has been a core theme in the discussions of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS), which has met formally since 2017 in the framework of the Convention on Certain Conventional Weapons (CCW).

The meaning and operationalization of control have been among the most contentious topics in the Group's debates. Three main modalities of control have emerged in recent years and are now widely considered to impose practical limits on AWS: control on the *weapon parameters*, control on the *environment of use*, and control through *human–machine interaction during use*.

Human–machine interfaces, which are the physical nexus between operators and the AWS, play a critical role in human control, and their role has been highlighted both at the GGE on LAWS and in various national policy documents. Human–machine interfaces are important to the development and retention of situational awareness, and to the architecture of control: allowing operators to monitor a system, and if necessary, to deactivate or to override it.

This report highlights several important dimensions of **the role of interfaces in human control over AWS**. It focuses in particular on the challenges, both present and anticipated, brought about by an increasing use of artificial intelligence (AI) and machine learning (ML) in such systems. The report draws, in several cases, on examples from automation in the vehicles industry, which can provide significant lessons in terms of controllability and system design.

Key findings and conclusions of this study are as follows:

- It is important, first, to situate the discussion about the role of the interface in the **context of human–machine interaction** in autonomous systems, which imposes significant performance requirements for human operators, and which comes with inherent risks such as over-trust or under-trust in the technology, which are further exacerbated by the use of AI/ML.
- For an interface to be an effective means of control, it must have a **high degree of usability** (meaning it must be engineered and developed in a way that enables the users to achieve their goals), and **operators must be adequately trained** to use it effectively. The achievement of these criteria entails important changes with the introduction of AI and ML in weapon systems: as weapon systems become more complex (e.g., endowed with more autonomy in critical functions), interfaces are poised to become more complex, as are the training requirements for human operators.
- **Recent research in interface** design indicates a focus on the interaction between humans and AI and on '**human–AI teaming**', and how that must be reflected in the design process. This is supported by the belief that as machines become more complex, options for human–machine *interaction* must evolve accordingly.
- **Personnel training** in the context of systems that continue to learn over time prompts the need for additional training curricula and methodologies that can support operators in building appropriate mental models of the systems, and in calibrating trust and expectations.
- One way to **mitigate issues of trust and incomprehensibility** in systems reliant on AI/ML is to embed more options for **explainability and transparency of the AI process** into interfaces, such as with visualization techniques (e.g., dashboards that reveal part of the process or conclusions of the AI). These efforts are important but bring, themselves, additional complexities to human–machine interaction, which can compromise human control.


Sector 53

Introduction

Human–machine interfaces are the physical nexus between human operators and autonomous systems, and a critical element in the array of options for human control over systems. An interface combines both hardware and software, and can include a range of components, such as physical control panels with buttons, dashboards and touchscreens. Interfaces allow the human operator to monitor a process (e.g., navigation), to modify or configure control settings, to adjust parameters and commands, or to manually override the system's operation. They can also display critical information and present the operator with an understanding of both the system's status and, in the case of remotely operated systems, of the environment in which that system operates.

Interfaces have featured in the discussions of the Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapons Systems (GGE on LAWS) on numerous occasions during the Group's deliberations on **human control over autonomous weapon systems** (AWS). Understandably, interfaces have been considered important because they provide at least two critical means of retaining a degree of control: allowing operators **to monitor** the behaviour and actions of a system, and **to deactivate or to override** it (e.g., by manually taking control) should it fail to perform as expected. As systems become more autonomous, however, interfaces become more complex as well.

This report analyses the role of interfaces in the exercise of human control. It presents several aspects of interface design and use in the context of AWS and highlights important trends on the horizon as more AI-enabled functions are to be incorporated in weapon systems.

Generally, for an interface to be an effective means of control, it must have a **high degree of usability** (which means it must be engineered and developed in a way that enables users to achieve their objectives) and operators must be adequately **trained** to use it effectively. The realization of these fundamental criteria becomes more complicated with the scaling up of autonomy and the use of AI and ML in weapon systems—as weapon systems become more complex, options for human–machine interaction and for interface design become more complex, as are the training requirements for human operators.

Because interfaces cannot be meaningfully discussed as stand-alone capabilities, the report integrates this analysis into the broader context of **autonomy**[1] **and human–machine interaction**. The challenges brought about by autonomy in weapons systems are weaved into all aspects of interface design and use, and in personnel training requirements.

1 **Note on terminology**: this report refers to the concepts of 'autonomy' and 'automation'. Automation generally refers to systems that are deterministic and predictable, whereas autonomy comes with less predictability (depending on the level of autonomy) and less deterministic behaviour. To date, the vast majority of available research is in automation because this field has existed for much longer, hence the numerous references to automated systems and automation in this report. From an end-user perspective, however, the distinction between automated and autonomous systems is not always visible and the issues faced may be very similar. As a forerunner to autonomy, automation provides important lessons to take into account and which can inform the policy community's evaluations of risks and challenges in autonomous systems. Further, the distinction between autonomy and automation can also be characterized in terms of the amount of human control and a system's ability to operate without human interference. In this sense, a weapon system that uses machine learning or deep learning to detect and define the type of target, and then presents that information to the human operator who can decide to engage or not, is not autonomous per se. Rather, such a system incorporates functions that have varying degrees of autonomy.

The report begins with an overview of the characterization of human control over AWS in multilateral and policy discussions, including in the GGE on LAWS, and provides a general introduction to the role of interfaces in human control (**Section 1**). It then situates the discussion of the role of interfaces in the broader context of human–machine interaction, which highlights key challenges for human operator performance (**Section 2**). The subsequent section presents the main approaches to interface design. The description of various approaches hints at important considerations for human control: system design has evolved to increase usability, but the introduction of AI/ML is also rendering the technology that supports human–machine interaction more complex (**Section 3**). In addition to system design, training is another critical factor in considerations of human control. More autonomy and more complex interfaces introduce new challenges for training for men and women and all members of armed forces, especially in the context of AI/ML-enabled systems (**Section 4**). Handling the complexity of systems that use AI, and which continue to learn and adapt over time, is difficult and the lack of predictability and transparency in these systems can impact trust in and reliance on the technology. Efforts to address this complexity with approaches to explainability and transparency ('Explainable AI', or 'XAI'), such as through visualization techniques, are important but continue to present limitations **(Section 5**).

1. Human–machine interfaces and human control

High levels of autonomy in a system may conjure up images of machines that perform actions alone but in reality, no system to date—no matter how autonomous—is entirely independent of *some* form of human control or supervision.[2] As the goal of full autonomy in weapon systems is neither feasible at the moment, nor desirable from a military efficiency standpoint,[3] human–machine interaction remains central to discussions of autonomous weapons systems.

Human–machine interaction has been a core theme in the discussions of the GGE on LAWS. The importance of this topic was reflected in the Group's Guiding Principles, adopted by consensus in 2019, particularly Principle C, which states:

> (c) Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole.[4]

Human control emerged therein as a critical concept and has remained central to the Group's debates over the years.[5] Despite many divergences about which capabilities need to be banned outright or not, there is broad consensus among State Parties that, regardless of the degree of autonomy in a weapon system, a certain level of human control over autonomous weapon systems needs to be maintained.[6]

Several critical themes related to autonomy in warfare (e.g., responsibility and accountability, applicability of international humanitarian law) are effectively within the purview of the concept of 'human control'. This echoes a similar development in the use of AI, ML and robotics in the civilian sector, where discussions around 'controllability' have taken center stage, combining "complex technical, ergonomic, legal, moral and organizational factors".[7]

1.1. Practical measures for human control and the role of interfaces

Distilling the meaning and parameters of human control in the context of AWS has been challenging, complicated further by differences among systems and operational environments.[8] However, general tangible measures of human control that have been proposed encompass considerations of system design and use.[9] This taxonomy of control illustrates two different but complementary modalities for exercising human control, both through the **design** of the weapon system itself, which includes both hardware and software elements, and operational control during **use**.

2 Endsley (2017).
3 See GGE on LAWS (2018a); GGE on LAWS (2018b).
4 GGE on LAWS (2019b)
5 The concept of 'meaningful human control', introduced by Article 36, has been highly influential as it refers more specifically to human control over *critical decisions in the use of lethal force*. It therefore goes further in providing more precision (referring to 'control' rather than the 'loop' or human 'judgment') and qualifying the nature of that control (i.e., 'meaningful'); UNIDIR (2014, 3). The theoretical, practical and legal ambiguities of the term were however not resolved (e.g., what *legal regulations* derive from this principle?). See Santoni de Sio and van den Hoven (2018). This report will refer to the broader term of 'human control' because it has been more generally used in technical literature.
6 Schwarz (2021).
7 Boardman and Butcher (2019, 2).
8 Schwarz (2021) notes that the discussion on control is misleading when discussing systems that increasingly take on decision-making roles, whereas forums such as the GGE on LAWS continue to embrace an instrumentalist position on technology, which assumes that technology is a tool over which their users retain full agency. A similar question is raised in Section 4.2. in the context of the rejection by the GGE on LAWS of anthropomorphic language (i.e., Principle (i) of the Guiding Principles).
9 iPRAW (2019).

Box 1. *Practical measures of human control over AWS*

The 2020 SIPRI and ICRC Report "Limits on Autonomy in Weapon Systems"[10] proposed **three practical measures to exercise control**: 1) control of the weapon parameters (such as, type of target); 2) control of the environment of use (for example, by limiting the use of AWS to specific locations/areas); and 3) control through human–machine interaction during use (such as by retaining the ability to supervise an AWS). A similar classification has featured in the discussions of the GGE on LAWS, including in 2020, when the Chair's summary lists these same three elements of control as a basis for engaging in further deliberations:

> States shall ensure a human operator or commander exercises judgement over the operational context, including through constraints on, inter alia, tasks, target profiles, time-frame of operation, and scope of movement over an area and operating environment, applied in individual attacks; in other words, constraints applied to the weapon system, the parameters of the weapon system's use and the required interaction between human and weapon system.[11]

The role of **human–machine interfaces** (HMIs) has been highlighted on several occasions as critical to the operationalization of human control, with HMIs spanning both criteria of system design and use.

The role of interfaces is particularly important in the operation of uncrewed systems controlled from a remote (sometimes, very distant) location, where a central feature of the system control is that operators' sensory connection with the machine is mediated by an interface.[12]

Rendering an autonomous system controllable does not hinge on interfaces alone, nor is it limited to the ability of the human operator to take manual control of the machine.[13] However, as the nexus between humans and machines, interfaces are critical to the control of an AWS, with direct consequences for how lawfully a system is used. While parameters of control may be exercised in multiple ways, and 'distributed' across the system's design (e.g., types of targets etc.), interfaces afford operators the possibility to monitor the system and to intervene should other forms of control falter or when circumstances on the ground change and "invalidate planning assumptions".[14]

1.2. Interfaces and the architecture of autonomy

User interfaces are defined in ISO standard ISO 9241-110:2020 as the "**set of all the components of an interactive system that provide information and controls for the user to accomplish specific tasks with the interactive system"**.

10 Boulanin et al. (2020, 8–9).
11 GGE on LAWS (2021, 6).
12 Riley et al. (2017, 180).
13 The Report of the GGE on LAWS from 2020 explicitly articulated this point: "Effective human control, involvement or judgment may not necessarily equate to direct, manual control but rather contextual factors including boundaries placed on the weapon and environment of use, and requirements for human-machine interaction" (GGE on LAWS, 2021, 8).
14 Boulanin et al. (2020, 9).

Interfaces are subsystems of human–machine systems[15] and "the window (both metaphorically and literally) through which operators interact with the machine".[16] Interfaces include numerous components, which vary depending on the system, such as input controls (for example, buttons and checkboxes), navigational components, information components, etc.

Generally, HMIs facilitate both input and output: input allows the operator to enter information into the technical system and output indicates the effects resulting from the input.[17] For example, one way in which the input–output loop may look for an AWS would entail the operator going through the interface menu and introducing certain coordinates into the system when engaging a target (input). The system would respond by providing its own coordinates and assessments, such as on collateral damage or other feedback based on operational procedures (output).[18]

1.2.1. Situation awareness

Human–machine interfaces play an important role in the development of **situation awareness (SA)**, which is central to human–machine interaction especially in dynamic environments. SA is impacted by both individual and organizational factors (such as stress, workload, task-switching requirements, or team dynamics) and by system factors, such as the mechanics of the system (e.g., the capacity of sensor technology to gather relevant data) and system interfaces.[19] The quality of the interface design can significantly and directly improve SA.

A sufficient level of SA will also help the operator realize, for example, that a certain situation is outside the bounds of automation capabilities, or that the automation is performing incorrectly.[20]

Box 2. *SA Levels 1-2-3*

The standard conceptualization of SA includes three hierarchical phases: "the *perception* of the elements in the environment within a volume of time and space, the *comprehension* of their meaning and the *projection* of their status in the near future".[21]

Level 1 SA. Perception of the Elements in the Environment: for example, for a tactical commander it means perceiving basic data on location, type, and capabilities of enemy and friendly forces in a given area.[22]

Level 2 SA. Comprehension of the Current Situation: for example, for a tactical commander, this means understanding how the appearance of a certain number of an enemy aircraft in an area of interest relates to or clashes with their objectives.[23]

Level 3 SA. Projection of Future Status: for example, a commander would be able to project that the current presence and offensive actions of a certain aircraft will lead it to attack in a certain way and certain area. This allows deciding on a course of action to achieve goals.[24]

15 The term 'system' here is used in the sense employed in the ergonomics domain, which refers to a system in the context of 'human–machine systems', describing various elements and the interaction between them. 'Weapon systems' are typically defined as a combination of weapons, related equipment, personnel, means of delivery, etc. In other parts of the report, however, the reference to systems is narrower (such as a computerized system).
16 Hou et al. (2015, 33).
17 Ibid.
18 Interview with Ming Hou (26 April 2022).
19 Endsley (2015, 11–12)
20 Endsley (2017, 8–9).
21 Endsley (1995, 36).
22 Ibid.; Gillan et al. (2017, 57).
23 Endsley (1995, 37).
24 Ibid. SA can be studied both at the individual level and at the team level. Every team member must have SA for the responsibilities assigned to them, and as a team they can develop *shared SA* when the goals of two or more team members overlap.

1.2.2. Understanding AWS status and behaviour

Interfaces are critical in representing key characteristics for control of an autonomous system, including system *observability*, *predictability* and *directability*.[25]

- Observability refers to the ability to observe and monitor the status of the system.
- Predictability refers to the property of understanding how the system behaves.
- Directability concerns the ability to influence the system.

While the technical literature includes different taxonomies of elements of control,[26] the goal of observability–predictability–directability synthesizes the fundamental requirements both for the AWS' autonomous capabilities and for interface design.[27]

25 Johnson et al. (2014, 9–10).
26 See, for example, Siebert et al. (2022).
27 Johnson et al. (2014, 9); interview with Matthew Johnson (31 March 2022).

2. Interfaces of AWS and human–machine interaction

Interfaces are not a stand-alone capability and their role in human control can only be understood in the framework of human–machine interaction. This section looks in particular at exigencies of human performance in the broader context of automation and autonomy, and the importance of framing the role of interfaces in relation to the context of their use.

2.1. New performance demands and challenges of cognitive involvement

The introduction of autonomy presents important changes in performance demands.[28] Developing effective systems is not only a matter of engineering and technical advances. Rather, "the most problematic aspect of … autonomous operations is the human aspect, or human–machine integration".[29]

Even when the human operator is 'merely' tasked with the oversight of an autonomous system, they face numerous challenges that may arise from data overload, from incomprehensibility of the system, from inadequate training, from interfaces that are not designed with the user's real needs in mind, and so forth.

For example, in the USS Vincennes incident of 1988, when the Iran Air Flight 655 commercial aircraft was shot down by an Aegis combat system stationed on the warship, a poorly designed weapon control computer interface led the plane to be incorrectly identified as a fighter jet. The display of information was overly complex and inadequate, giving the controllers the impression that the airliner was descending towards the ship, while in fact it was moving away from the ship.[30] Further, overconfidence in the Aegis technology ultimately led to 'overtrust' in the system, and a failure to challenge the system's identification.[31]

To mitigate some of the risks in human–machine interactions in military settings, it has been suggested that interfaces need to maintain the user's **cognitive involvement**, a point that has been embraced by several delegations at the GGE on LAWS.[32]

Maintaining **human cognitive involvement** in autonomous systems is challenging for at least two reasons.[33]

28 Hawley et al. (2005); Development, Concepts and Doctrine Centre (2018). Parasuraman and V. Riley (1997, 231) made a significant contribution to the discussion of how automation changes work for humans. The 'ironies of automation' have been, however, known for over four decades, following the introduction of automation in industrial processes and positing that the more advanced the automation, the more complex and crucial the contribution of the human. Bainbridge (1983); National Academies of Sciences, Engineering and Medicine (2022, 44).

29 Hawley (2017, 11).

30 Cummings (2006, 23); Scharre, (2018, 169–170).

31 Swartz (2001).

32 For example, in the Commentary on the Guiding Principles, Switzerland proposed that one possible way of exercising control could be achieved by "maintaining the ability of human supervision, by using technology (for instance appropriate human-machine interfaces) **to support the human cognitive involvement**". GGE on LAWS (2021, 88).

33 These points are further elaborated in the following sections, both in discussions of system design and personnel training.

1. The first is that, when assigned **supervisory roles**, humans are simply not able to retain attention constantly and consistently, and the expectation of sustained operator vigilance, which is needed for intervention at the right moment, is unreasonable.[34] Delegating a rather passive role[35] to human operators runs the real risk of getting them disengaged, which makes it difficult to maintain alertness. As one expert explained, "often in [command and control] centres, nothing is happening and then, all of a sudden something happens, and you need to get back into the cognitive loop. It is difficult to make sense of the situation, no matter how good the interface is".[36]

Some solutions for maintaining higher rates of vigilance have been proposed through interface design. A report of the Development, Concepts and Doctrine Centre of the UK Ministry of Defence suggests that interfaces can be optimized to support this goal, including by requiring operators 1) to **search** for defined objects (which is shown to enhance mental engagement), and 2) to **explore** things of interests, such as boundaries or anomalies.[37]

Other smaller and more subtle kinds of tasks can include features such as text messages and alerts prompting the operator to check system status.[38] Ultimately however the answer to maintaining vigilance is for operators to be engaged in meaningful tasks, "and not going to the back-end".[39] While training curricula play an important role in building supervisory skills, the allocation of (meaningful) tasks to human operators remains key to maintaining cognitive involvement.

2. The second reason can be explained by inherent challenges in human–machine interaction, including challenges of "automation complacency",[40] loss of attention that can occur as certain tasks get automated and attention goes to other tasks,[41] or ambiguous or inexact expectations from the system. Some of these challenges stem from what is known as the **'automation conundrum'**. This posits that the loss of human alertness is directly proportional to the system's enhanced automation and reliability: "The more automation is added to a system, and the more reliable and robust that automation is, the less likely that human operators overseeing the automation will be aware of critical information and able to take over manual control when needed".[42]

When a system that is highly automated and highly reliable fails, it introduces complicated performance challenges for the operator. Because high levels of automation increase dependence on the system, they simultaneously increase the likelihood of failed manual recovery.[43] This has been described as the **'lumberjack effect'**, exposing the tradeoffs that come from benefits of high reliability and the attendant costs of failure, similarly to trees in a forest: "the higher they are, the farther they fall".[44] In automation research,

34 Boulanin et al. (2020, 19); Development, Concepts and Doctrine Centre (2018); Hawley (2017, 9).
35 Challenges related to 'passive cognition' have been observed even with experienced air traffic controllers; see Endsley (2017); Metzger and Parasuraman (2001).
36 Interview with anonymous expert (25 March 2022).
37 Development, Concepts and Doctrine Centre (2018, 32).
38 Interview with Ming Hou (26 April 2022).
39 Interview with Mica Endsley (15 March 2022).
40 Parasuraman and V. Riley (1997).
41 Interview with Mica Endsley (15 March 2022).
42 Endsley (2017, 8).
43 National Academies of Sciences, Engineering, and Medicine (2022, 42).
44 Onnasch et al. (2014, 477). The metaphor of the 'lumberjack effect' has been widely used in human–systems integration research.

this situation has also been discussed for over two decades as a risk of **skill degradation**,[45] which concerns especially high-performing automated systems that function properly for a long time prior to the first failure. In such case, operators learn to rely extensively on the system, over-trust it, and even become complacent. Calibration of trust refers to the "correspondence between a person's trust in the automation and the automation's capabilities", and can manifest in either over-trust, or conversely, under-trust.[46]

Further, in the context of the use of AI, it is difficult to quantify "the ability of an AI system to appropriately calibrate and execute its expected functions".[47] The shift towards human–AI 'teaming' as a preferred paradigm in human–AI collaboration (elaborated in Sections 3.2. and 4.2.), and the notion that humans and AI need to collaborate as teammates with a shared goal, comes with its own sets of challenges.

2.2. Interface design and context of use

The design of an interface is critical for its usability[48] as it "can directly affect the operator's ability and desire to complete a task … to understand the current situation, make decisions, as well as supervise and provide high level commands to the robotic system".[49] For example, studies on UAV control for reconnaissance missions revealed that operators wanted to 'fly the camera', meaning that rather than devoting attention to controlling the vehicle and its systems, operators showed a strong preference for being able to position the camera where it needed to be to meet their mission objectives. Consequently, the design options were focused on user interfaces that would eliminate direct control of roll, pitch, and yaw.[50]

General principles of design for AWS have been addressed in various forums. For example, the United States Department of Defense Directive 3000.09 stipulates that "the interface between people and machines for autonomous and semi-autonomous weapon systems shall: a. Be readily understandable to trained operators; b. Provide traceable feedback on system status; c. Provide clear procedures for trained operators to activate and deactivate system functions".[51]

This recommendation was reiterated by the United States at the GGE on LAWS on several occasions.[52] The 2019 GGE Report lists "readily understandable human-machine interfaces and controls" as a possible risk mitigation measure, alongside measures such as "rigorous testing and evaluation of systems" and training of personnel.[53]

A great deal of focus on making interfaces 'clear' or 'readily understandable', however, risks being misconstrued as a need for simplicity. This may inadequately shift the attention to micro-ergonomics, or elements such as colour, font size, etc. in a display system, which although important are "not the place to start".[54] As a rule, display

45 Parasuraman (2000); Cummings (2006).
46 Lee and See (2004, 55).
47 National Academies of Sciences, Engineering and Medicine (2022, 18).
48 ISO 9241-220:2019 defines usability as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context".
49 Adams (2002).
50 Riley et al. (2017, 178–179).
51 US Department of Defense (2012, 2–3).
52 GGE on LAWS (2018c); US Mission (2021).
53 GGE on LAWS (2019a).
54 Interview with anonymous expert (25 March 2022).

functions need to be concerned with the control of 'mission-relevant information', while minimizing 'mission-irrelevant information'.[55] However, sense-making in complex systems is not limited to the data coming from a screen. The view that "technical and tactical complexity can be reduced to manageable levels … through 'proper' system and interface design"[56] is contested by many experts in cognitive systems engineering, who believe that this approach **does not reduce complexity but merely hides it from users**.[57]

This does not mean that the solution rests with *display complexity*,[58] but that other conditions of system interface design must be met for the interface to be conductive to enhanced human performance and control.

A fundamental requirement in the interaction with an AWS is to build a good mental model of the system prior to use (including understanding the system's levels of automation, and whether it is behaving appropriately and performing at the expected parameters), and to understand the system's scope of behaviour, as well as the system's likely changes over time—a key aspect of ML-based systems.[59] These requirements must be reflected in the design of the interface, as well as in different training needs.

55 Davis et al. (2017, 408).
56 Hawley and Mares (2017, 16).
57 Hollnagel and Woods (2005); Hawley and Mares (2017). A similar view was shared by several experts interviewed for the project.
58 See Endsley (2017).
59 Interview with anonymous expert (22 March 2022); interview with Mica Endsley (15 March 2022); interview with Mennatallah El-Assady (8 April 2022).

3. Approaches to HMI design

This section presents an overview of approaches to interface design in highly automated and autonomous systems, starting with 'human-centred design', followed by 'interaction-focused' approaches. As discussions of human–machine teaming[60] or human–AI teaming[61] proliferate, options for interface design become more complex.

Human–systems integration (HSI), a topic originating in the mid-1980s, addresses human considerations in system design and implementation. It is a 'total system' approach that aims for integration across systems, including humans, technology, operational contexts, and the interfaces among these elements.[62] As is shown in this section, the introduction of AI and ML (and a growing focus on human–AI teaming) has revealed limitations in traditional or cognitive systems-engineering approaches, and their inability to address *how* new systems need to adapt.[63]

3.1. Human-centred design

Two general requirements for interface designers are:

1. to understand what robotics operators across various tasks and domains need to know; and
2. to determine how to present the information in an integrated fashion in order to support situational awareness and decision-making.[64]

Integrating what the operators *need to know* into the process is fundamental to **human-centred design**. This paradigm emerged in the 1980s to address shortcomings in the technology-centred paradigm that was traditionally the standard and which meant that an interface would reflect first what the engineers creating the system considered essential, or what they viewed as relevant.[65]

Human-centred design, in contrast, is an approach to systems design and development that "aims to make interactive systems more usable by focusing on the use of the system; applying human factors, ergonomics and usability knowledge and techniques".[66] A human-centred design, whether it is conceptualized as soldier-, customer-, or user-centred, seeks to optimize the interface "around how people work, rather than force people to change how they work to accommodate the system".[67] It regards the operator as "a component of the system just like the sensors or underlying code", and whose capabilities must be incorporated into the design.[68]

60 The Development, Concepts and Doctrine Centre of the UK Ministry of Defence proposed the term 'human–machine teaming' in a 2018 Joint Concept Note and this position was reiterated in an Expert Paper submitted to the GGE LAWS in 2020, referring to human–machine teaming as "an approach which recognizes that the integration of humans and machines working towards a common goal, with their relative strengths and weaknesses, is key to military success". GGE on LAWS (2020, 2–3).
61 Endsley (2017, 6); National Academies of Sciences, Engineering and Medicine (2022).
62 See, for example, US Department of Defense (2022). HSI is a widely used concept and approach.
63 National Academy of Sciences, Engineering and Medicine (2022, 71).
64 Gillan et al. (2017, 57–58).
65 This does not mean that users' needs were previously sidelined in the design process. For example, the concept of **'participatory design'** was proposed in the 1970s to integrate the expectations and creativity of the user in the design process; see Flemisch et al. (2008). Design problems persisted, however, leading to many shortcomings in how information was integrated in displays as systems became more complex (Endsley, 2013); the integration of users' perspectives remained in practice often superficial (interview with Matthew Johnson, 31 March 2022).
66 ISO 9241-220:2019.
67 Savage-Knepshield (2017, 276).
68 Oury and Ritter (2021, 22).

In practice, human-centred design follows an **iterative process** that begins before the development of the interface[69] and requires integrating both design and evaluation through "incremental development and iterative refinement" of the system, based on input and feedback,[70] understanding the users,[71] and integrating their perspectives from the beginning.[72] During this process, there may be emerging properties in the system, or people may use the system in different ways. This will impact several choices in the design, including the amount of cognitive load on the operator, their reliance on the system, and the choice of when to use or when to turn off certain functions.[73] In short, it is a process of "evidence-based evolutionary tinkering".[74]

With an increase in autonomous functions, additional requirements of human-centred design have been suggested in order to support the operator in their understanding of the system's functionality. In addition to effectively presenting the needed **information for decision-making**, it has been suggested that interfaces must include **cues related to the state of the automation** (including modes and system boundary conditions), support for **mode transitions** (including, for example, the necessary support for transition to manual control), and system **transparency** that provides understandability and predictability of the system's actions.[75]

3.2. Interaction-centred approaches

More recent approaches to design of intelligent systems, beginning with the 2010s, focus on the design process as shaped by and responding to the interaction[76] and interdependence between humans and machines.

Strictly speaking, interaction was always part of the design process, only in different modalities, and the fundamental principles of human-centred design have not been abandoned. However, research over the past decade has focused more closely on capturing the collaborative dimension of human–machine interaction, and as technology now affords a more rapid adaptation to a system's learning. This evolution is underscored by an emerging understanding that building effective autonomous systems relies upon a successful approach to human–autonomy teaming,[77] or human–AI teaming and that as machine capabilities expand, human–machine *interaction capabilities*[78] must also expand.

For example, one approach, **"coactive design"**, assesses the design implications that follow from human–robot teaming, where both humans and systems participate *simultaneously* at completing a task, and where systems need to be designed to support coordination, collaboration and teamwork.[79] This design approach regards teaming as a process that involves both parties (humans and AI systems) and is premised on the *interdependence* that exists in the interaction between the two.

---

69 Interview with anonymous expert (8 April 2022).
70 Savage-Knepshield (2017, 275).
71 Suggested **methods for designers to learn about users** include a menu of actions, from simply talking to them, watching them work, having them use interfaces created by the same designers, to getting more general information about their work environments (Oury & Ritter, 2021, 16–17).
72 One expert added that what is important is to integrate the users' perspectives "at the *right* moment, not necessarily as early as possible" (interview with anonymous expert, 8 April 2022).
73 Interview with anonymous expert (22 March 2022).
74 Interview with anonymous expert (25 March 2022). The same expert noted that **organizational factors** matter too: the process must be supported by an organizational culture in which honest feedback is possible and taken into account. Boardman and Butcher (2019, 11) make a similar point in discussing dimensions of human control, which has an organizational dimension as well, insofar as the organizational culture must not impact the "willingness to question system behaviours and actions".
75 Endsley (2017, 10); Endsley, Bolte and Jones (2003).
76 Interaction is defined by ISO/TS 20282-2:2013 as "bidirectional information exchange between users and equipment". Equipment includes both hardware and software. Information exchange may include physical actions, resulting in sensory feedback.
77 Endsley (2017, 5).
78 Johnson et al. (2018).
79 Johnson et al. (2011).

Other models of cooperation in human–AI teams have been proposed in the form of **"dynamic task allocation"**,[80] meaning that the riskiest and most morally salient tasks can be allocated to humans, while the other decisions are assigned to artificial agents. This approach requires that explanations be intrinsically part of the human–agent collaboration (see section on explainability) and part of the interface design.

Another approach called **"co-adaptive guidance"**[81] is based on a similar principle that interfaces need to adapt based on feedback from the user and to calibrate cognitive involvement, trust and changing expectations over time. This approach accounts for the 'three moving parts' in human–systems interaction in systems that learn over time: 1) the changing mental models of the *human operators*, as 2) *facts* on the ground are also changing, and as 3) *the system/AI model* itself is changing.[82]

This would require the interface to synchronize and represent the system's learning and adaption. For this interaction to be effective, it also requires that the system become more of an agent in the sense that it is able to prompt the operator to intervene, such as by communicating to them 'I need input' or 'I do not know' in certain situations.[83] The co-adaptive *learning* that supports this approach, also entails that the system's model of the user will guide it to detect inconsistencies or contradictory signals and can thus function as a safeguard. For example, the system could stop altogether if the operator is suddenly replaced.[84]

Some of these principles are applied in the autonomous vehicles industry, where interface design is increasingly approached as a **"co-creation process"**, meaning it aspires to integrate drivers' preferences or to introduce corrective elements, such as by prompting a young driver to pay more attention.[85] In the autonomous vehicles industry, user-specific adaptability of interfaces is considered important for the future of the industry and for garnering more trust in the technology, especially at higher levels of automation.[86] Increased levels of automation however will require a tradeoff in data from the driver, including more biometric data,[87] because "as you give more control and build more trust, the vehicle needs to know more about you; now you as an operator need to be monitored".[88]

In the military, the use of **biometric and neurophysiological data** in interface design is not a new idea, although it remains for now largely exploratory. For example, simulations conducted in air traffic control have employed eye movement parameters to understand cognitive demands as well as fluctuations in cognitive workload following different kinds of displays (e.g., cluttered weather displays that complicate the pilot's effort to extract relevant data).[89] While this applies to testing and simulation protocols, the use of biometric and neurophysiological data could be integrated in a real-time closed-feedback loop analysis system. Such a loop would assess a user's interactions and update the system about the user's state and ongoing cognitive load.[90]

80 van der Waa et al. (2020); van der Waa et al. (2021).
81 Sperrle et al. (2020).
82 Interview with Mennatallah El-Assady (8 April 2022).
83 Idem.
84 Idem.
85 Interview with Parrish Hanna (16 March 2022).
86 Hartwich et al. (2021, 14–15).
87 Examples include gesture recognition, eye tracking, facial recognition, fingerprint and voice biometrics. Burt (2020).
88 Interview with Parrish Hanna (16 March 2022).
89 Ahlstrom and Friedman-Berg (2006, 623–624).
90 Hou et al. (2022, 11).

The technical scholarship that highlights the need for more 'user state' data in systems underlines the fact that, as systems acquire more decision-making capabilities, human–machine teaming depends on elements of trust from both parties. In the context of autonomous systems, the meaning of trust[91] refers to *verification mechanisms* embedded in the system. Such mechanisms would ensure, for example, that the operator's inputs are consistent and not impacted by stress.[92]

Finally, another design approach includes **immersive interfaces**, which have been researched in recent years in various domains, including automated vehicles[93] and uncrewed aerial vehicles.[94] Immersion and immersive technologies refer to virtual worlds that are simulated, dynamic, and include elements such as rich three-dimensional spaces and high-fidelity motion.[95]

Virtual reality (VR) and augmented reality (AR) systems have been used for military training and to create simulated environments. However, an immersive interface would be used for mission execution and in order to promote collaborative behaviour (i.e., human–machine collaboration).

Immersion, as a design choice, is considered a more natural collaboration platform and is promoted as a practical tool to visualize both the physical world and its mirrored visual presentation with the same level of dimensionality.[96] It has been suggested previously that the remoteness and distancing created through the interface introduces a 'moral buffer' that allows operators to distance themselves from their actions, and from negative consequences.[97] Immersion may also be a useful method to reduce cognitive and moral distancing, a concern raised repeatedly in the case of AWS[98] and, generally, about weapons systems operated from a distance, such as drones.[99]

91 In the deliberations of the GGE on LAWS, the use of **anthropomorphic language** (e.g., trust) attributed to LAWS was rejected as a matter of principle, including with Guiding Principle (i), which states that "In crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized". Arguably, the technical use of the term 'trust' in this case refers to a system component, i.e., a verification mechanism, and through which it would 'verify' the user's inputs, without impinging upon human responsibility and legal obligations. See, however, Box 3.
92 Hou (2020); Hou et al. (2022, 14–17).
93 Georg and Diermeyer (2019).
94 Feuerriegel et al. (2021).
95 Schultze and Orlikowski (2010).
96 Feuerriegel et al. (2021, 65).
97 Cummings (2006, 26). Cummings cautions against the use of elements in design that may further exacerbate the sense of remoteness and lack of responsibility, such as graphic elements that make the interface appear like a video game.
98 The 'cognitive distancing' introduced by AWS may be driven by both temporal and spatial distancing: there may be hours, days, weeks between activation and application of force (temporal distancing), and uncertainty about the location where the use of force will be applied (spatial distancing). Boulanin et al. (2020, 12).
99 Coeckelbergh (2013). The author makes the point that drone fighting (or 'screenfighting') introduces not only physical distance but also moral distance between the fighter and their opponent.

4. Challenges for training

Training of operators of AWS is an important element of control. For example, in the deliberations of the GGE on LAWS, characterizations of control refer to a weapon's entire life cycle, which also includes training (see Annex A).

Autonomous systems introduce new types of training requirements for human operators, which depend on the properties and complexity of the system interface.

For a start, operators need to understand the system more holistically, in terms of its *functional bounds and the functional allocation between human and machine*.[100] Further, training in the context of autonomous systems that use AI is challenging as the systems evolve and keep learning.[101] Training requirements become more complex compared to static systems.[102] The way that systems change their internal models is often opaque and hard to understand even for their developers[103] and it is difficult to present a model of learning that an operator can train on because the system will learn differently in different environments (e.g., training phase vs. operational environment).[104]

4.1. Understanding levels of autonomy and functional allocation

The distribution of tasks and the relationship between *perceived/attributed level of autonomy* and the *actual level of capacity* of the autonomous system is critical in calibrating trust and reliance on the system.[105] Preliminary studies in the context of autonomous vehicles reveal important lessons about the inherent risks in transitions between levels or modes of automation, and when the human operator (driver, in this case) does not accurately assess the system's limitations. For example, a particularly vulnerable area has been identified at the point between partially and highly automated systems, which leads to crashes when the driver perceives the car to be more automated than it is at a given moment.[106]

Challenges that arise from an insufficient understanding of a system's real capabilities have also surfaced in the military, for example, with the Patriot fratricide incident of 2003, where a US Army Patriot missile system shot down a UK Tornado and a US Navy F/A-18. The displays of the system were confusing and at times incorrect. The operators had 10 seconds to veto a computer solution and lacked training "in a highly complex management-by-exception system".[107]

---

100 GGE on LAWS (2020); Janssen et al. (2019, 101); interview with anonymous expert (22 March 2022). Function and task allocation between humans and machines is an 'evergreen' theme in human–automation interaction (Janssen et al. (2019, 101).

101 Interview with Mica Endsley (15 March 2022); interview with anonymous expert (22 March 2022); interview with Mennatallah El-Assady (8 April 2022).

102 A system that is 'static' may also use AI, but learning takes place in known, observable and deterministic environments. Such a system may employ, for example, a **search algorithm** to construct sequences of actions to achieve a goal, or problem-solving algorithms, which could be used for **planning**. Russell & Norvig (2022, chp. 3, chp. 14). Modelling dynamic situations, and situations that present degrees of uncertainty over time introduces more challenges. Russell and Norvig (2022, chp. 15).

103 Holland Michel (2020).

104 Interview with Mica Endsley (15 March 2022).

105 Flemisch et al. (2017); see Bahner et al. (2008).

106 Flemisch et al. (2017, 323–324). This was called the 'unsafe valley of automation', an expression that draws on the metaphor of the 'uncanny valley'. The authors of the study did not conclude that automated systems, or higher levels of automation are unsafe per se, but that "there are unsafe regions around safe automation designs and combination of different assistance and automation levels of transitions between levels or modes" (327).

107 Cummings (2006, 23); see also Scharre (2018, 137–145).

Experience with such systems led an engineering psychologist with the US Army Research Laboratory to conclude that **"an automated system in the hands of an inadequately trained crew is a de facto fully automated system"**.[108] Inadequate training can lead to incorrect expectations, an inability to cope with system failures,[109] or an inability to override the system's course of action, rendering it effectively "fully autonomous by neglect".[110] Interface design alone cannot compensate for highly effective training.

Training in the context of autonomous systems must focus on developing **operator expertise**.[111] This includes both quantitative and qualitative elements:[112]

- **qualitative**, with a more rigorous focus on the development of mental models of the system, and with a view to ensure that training is not merely 'habit transfer' (a common challenge when using a new interface because operators will tend to refer to older models).[113] This includes understanding, for example, variables such as the extent of autonomous functions, the system's changes from one environment to another, as well as where the system is most vulnerable or its uncertainties;[114] and
- **quantitative**, including changes to duration of training,[115] or, for example, to the intervals for updating the training.[116]

It is however important to note that there remain persistent human limitations to training in the context of supervisory roles. **Sustained vigilance** in supervisory roles is recognized to be a very difficult task.[117] Keeping focus is a matter of both selection and training, and it is now well known that some people are better at sustained vigilance than others.[118] However, even with appropriate selection processes and mandatory training, longer times spent in repetitive or supervisory tasks will lead to an increase in error rates. A solution for keeping operators cognitively involved hinges on more complex factors including on an incremental use of autonomous functions[119] that would allow human operators to better understand the system, learn when and how to revert to manual control, and to avoid losing the sense of responsibility.[120]

4.2. Training for human–AI teams

Finally, in the context of increased interdependence between humans and AI systems, an increasing body of research underlines that training must be adapted to take into account **human–AI teaming and teamwork**.

108 Hawley (2017, 9).
109 Janssen et al. (2019, 101–102).
110 Interview with Matthew Johnson (31 March 2022).
111 Hawley (2007, 9). Hawley refers to automated systems, but this taxonomy largely holds for autonomous systems.
112 Ibid., p. 11.
113 Interview with anonymous expert (25 March 2022).
114 Interview with Mennatallah El-Assady (8 April 2022); interview with anonymous expert (8 April 2022), who also noted that this is **not just a technical problem, but also a cultural one** as engineers often tend to highlight where the system works best but have difficulties being transparent about the system's weaknesses.
115 See Hoffman et al. (2014, 13); and Hoffman et al. (2009) for the conceptualization of 'accelerated expertise' and training for complex systems and for highly dynamic environments.
116 Interview with Mica Endsley (15 March 2022).
117 Hawley (2017, 9); interview with anonymous expert (22 March 2022); Endsley and Kiris (1995).
118 Interview with anonymous expert (25 March 2022).
119 Interview with Matthew Johnson (31 March 2022).
120 A US Air Force Report ("Human–Autonomy Teaming") suggests the concept of 'flexible autonomy', which stipulates that levels of autonomy can shift over time, 'back and forth' between airmen and autonomous systems, either at human discretion, or based on criteria built into the autonomy (such as when the airman loses connection with the system, or when there is not enough available time for in-the-loop control). United States Air Force (2015, 9–12); National Academies of Science, Engineering, and Medicine (2022, 45).

This is prompted by the expectation that with increased levels of autonomy in weapon systems, humans and AI systems will coordinate to perform high-complexity tasks as an integrated unit. Training in this context cannot be limited to transfer of knowledge and it will increasingly entail training together. This implies that the two sides will interact as 'peers', each contributing their own expertise and authority to take action.[121] In this case, the aim of human–AI training will need to be focused on *working* together and *learning* about one another.[122]

This entails two major shifts in training compared to human–human team training:[123]

- **perceptual** changes, which refer to issues such as bias,[124] trust and verifiability (an inherent challenge for AI systems), which humans expect and demand of AI systems, and which can lead to negative bias towards AI;[125] and
- **procedural** shifts, which include new methods of taskwork and teamwork training, including a need for designing appropriate simulation-based training in both live and synthetic environments.[126]

This brings additional requirements for the human operator who needs to understand 1) their role; 2) the AI system; 3) how to interact with the AI system/teammate; and 4) how to interact with the other human teammates. [127]

Research in this field is in its early stages but will be critical if and when existing approaches to teaming and training cannot support the full scale of complexity brought by the introduction of more autonomous functions.

121 McNeese et al. (2021, 3).
122 Interview with Matthew Johnson (31 March 2022). Johnson likens such training to that conducted in **surgical teams** that train to learn close coordination in uniquely demanding environments.
123 National Academies of Science, Engineering, and Medicine (2022, 63).
124 Chandler (2021).
125 Zhang et al. (2020, 4–5).
126 Development, Concepts and Doctrine Centre (2018, 47).
127 National Academies of Science, Engineering, and Medicine (2022, 67).

Box 3. *'Human–AI teaming', anthropomorphism, and Guiding Principle (i)*

The use of concepts such as 'peers' or 'teaming' in relation to AI systems (which evokes **anthropomorphism**) in technical literature does not situate innate systems on par with humans in terms of accountability and legal responsibility. Rather, the purpose is to underline the fact that with more autonomy comes more complexity and interdependence in human–machine interaction, and that training together is the only way to understand the system, build trust, achieve effective human–systems integration, and calibrate expectations. The SIPRI/ICRC Report "Limits on Autonomy in Weapons Systems" makes the point that the insistence on the concept of human–machine teaming in strategic documents indicates that the military wants to ensure that humans continue to exert agency and control over AWS.[128]

However, this arguably remains a point of consideration for the GGE on LAWS, which does not promote the use of anthropomorphic language (See **Principle (i)** of the Guiding Principles) in crafting policy measures for weapons systems based on emerging technologies in the area of LAWS. Even as the technical community employs words such as 'teaming' as metaphors (rather than assigning agency to autonomous systems), it is important to consider how **the technical usage of anthropomorphic language is not interpreted in a way that would violate this principle**.

The GGE could benefit from further elaboration on:

1. whether it needs to further qualify and elaborate on the principle in light of the framing of human–machine interaction emerging from technical scholarship; and
2. how to ensure that antropomorphic language is not misinterpreted and that it does not complicate understandings of legal responsibility and accountability.

128 Boulanin et al. (2020, 17).


N
0.2957
30°
75°

5. AI explainability and transparency

The 'black box' nature of intelligent systems[129] complicates the interaction with the end user and can result in inaccurate mental models, creating either too little or too much trust.[130] Efforts to make AI more transparent and explainable have proliferated in recent years as it has become increasingly recognized that AI's opacity negatively impacts trust in the system and its decision-making mechanisms.

An explanation system embedded in an interface could theoretically mitigate some of these risks during the use of an AWS but there are many challenges (and even potential drawbacks) with the existing available methods.

Explainable AI (XAI) is the field focused on the understanding and interpretation of (the behaviour of) AI systems.[131] **Explainability** is different from **transparency** as explainability evaluates the system's processes in a *backward-looking manner*, meaning it is looking into *what the machine did* and provides post hoc explanations. Display transparency, in contrast, provides *real-time understanding of the system's actions*. While in a military operational setting, transparency is arguably more valuable in supporting decision-making in real time, both explainability and transparency are important in building SA. Explainability, when time permits, can improve *review processes* and the mental model of a system, which can impact future SA.[132]

Box 4. *Explainability and interpretability*

Explainability and interpretability are very closely related, and often used interchangeably, although the two concepts are different.

Interpretability is the ability to present outputs in terms that are understandable to a human.[133] It refers to the quality of a system to provide enough data for a human to be able to predict an outcome.

Explainability refers to "the internal logic and mechanisms that are inside a machine learning system"[134] and the ability to explain those mechanisms in human terms.

An interpretable model means that the input-output relationship can be formally determined but it does not necessarily entail that humans can understand its underlying processes.[135] This subtle difference reflects a question in XAI research of whether a model of explanation should be aligned to human understanding or to the machine's model.[136]

129 'Black box' models refer particularly to machine learning and deep learning. There are also 'white box' or 'glass box' models, which produce more easily explainable results (such as linear models) but they are far less powerful compared to black box models. The trade-off between high performance and the model's "ability to produce explainable and interpretable predictions" continues to define AI. Linardatos et al. (2021, 1).
130 Sartori and Theodorou (2022, 4).
131 Linardatos et al. (2021, 2).
132 National Academies of Science, Engineering, and Medicine (2022, 33).
133 Doshi-Velez and Kim (2017, 2).
134 Linardatos et al. (2021, 3).
135 See Gilpin et al. (2019); Mueller et al. (2019, 85).
136 Interview with Mennatallah El-Assady (8 April 2022).

Box 5. *LIME (Local Interpretable Model-Agnostic Explanations)*

LIME is one of the most popular interpretability techniques for black box systems. LIME provides *local interpretability* by perturbing one local dataset (such as by tweaking values) and observing how the output changes. The output of LIME comes in the form of a list of explanations reflecting how each feature contributed to the prediction.[137]

5.1. XAI dashboards and their limitations

Most approaches to XAI have concentrated on visualization techniques through interfaces and dashboards that display parts of the AI process. For example, explanation interfaces can range from **dialogue boxes** and **graphical representations** in the form of a pie chart that shows probability, to interactive interfaces where users can interact with the system by selecting the best algorithm out of several.[138]

Visualization can help to foster more trust in the AI system, as well as more human agency.[139] Studies have shown, for example, that providing information related to the system's uncertainty improved performance, including performance of human take-over from the system.[140]

Box 6. *Criteria of XAI*

An underlying challenge in XAI is that it is often unclear what criteria of 'explanation' and 'explainability' it rests upon. In order words, what does it mean for a system to be explainable? What exactly should an explanation be about? Or what is the benchmark for assessing an explanation is optimal or satisfactory enough?

Evaluating measurement standards for XAI is the subject of ongoing research. Examples of some groupings of measurement criteria include:

- 'explanation goodness', which evaluates properties and elements of an explanation that make for a good explanation—they should be complete, logical, incremental, 'non-overwhelming', etc.;[141]
- performance improvement, which analyses measurements as to what extent the explanation enables the operator to use the AI in their work to achieve their objectives or to make predictions; [142] and
- impact on user's understanding/mental model of a system.

137 Hulstaert (2018). It is important to note that LIME has been shown to be at times unreliable.
138 Mueller at al. (2019) provide an extensive literature review of XAI studies and interfaces.
139 Beauxis-Aussalet et al. (2021, 10).
140 National Academies of Science, Engineering, and Medicine (2022, 35).
141 Mueller et al. (2020); Mueller et al. (2019, 99).
142 Mueller (2019, 95).

While the 'XAI Pipeline' includes a multitude of ML models and visual analytics methods,[143] they still present a high degree of complexity, which makes them mostly comprehensible to ML experts only.[144] Other limitations of XAI include that:

- **trust** is not a solely technical problem—it is a dynamic process, and visualization cannot address all problems related to trust;[145]
- **explanations** can reinforce flawed mental models, they can overwhelm people with details or include too many loose ends,[146] they can be persuasive tools in cases where further verification may be needed (and thus lead to situations of over-trust), and they can be interpreted differently by different users;[147] and
- in combat situations, some explainability tools may in fact **increase workload** in high-intensity operations, such as when an operator has a limited amount of time to review a system's explanation.[148]

5.2. XAI and autonomous weapon systems

While more transparent and explainable systems should be preferred to black boxes, it is important to bear in mind that explainability is not a silver bullet for enhancing trust and poorly implemented methods can be counterproductive. In the case of AWS, much more research is needed to define the best methods for system transparency and types of transparency information, including information across classes of operations.[149] This becomes clear, for example, when considering some options proposed previously which would have interfaces display elements, such as percentages of probability (e.g., 87% probability that X is a legitimate target) to the user as a method to enhance trust in the system. Such an approach can add, in fact, further difficulties to the human operator's decision-making process, and does not amount to much else than an "AI that mimics a relationship with the human via numbers",[150] without meaningfully reassuring the operator. Is 87% probability enough to proceed and engage a target? What does the remaining 13% mean? Furthermore, should the information be presented as '87% certainty' or as '13% uncertainty'?

Other dilemmas arise too in the context of a hypothetically fully transparent and explainable AI, which could mean that now the onus in case of error falls entirely on the human operator. This scenario risks introducing a disproportionate amount of responsibility and accountability for the human operator(s), given that AI cannot be held accountable because no such mechanisms exist.[151]

The future development of XAI methods, though not without promise, need to be carefully crafted and integrated into systems so as to effectively facilitate human–machine interaction in order to enhance human trust and human control.

143 Spinner et al. (2019).
144 Interview with Mennatallah El-Assady (8 April 2022).
145 Beauxis-Aussalet et al. (2021, 7–8).
146 Hoffman et al. (2018, 16).
147 van der Waa et al. (2021, 4).
148 Holland Michel (2020, 17); see also Kunze et al. (2019); Poursabzi-Sangdeh et al. (2021).
149 National Academies of Science, Engineering, and Medicine (2022, 35).
150 Interview with Matthew Johnson (31 March 2022). A similar point was made by Mica Endsley (interview 15 March 2022).
151 Interview with Mennatallah El-Assady (8 April 2022).

```
R.id.submitButton: {
tring url = "http://thebear.ef

ist<NameValuePair> params = ne
f(MaleRadio.isChecked())
    params.add(new BasicNameVal
lse if(FemaleRadio.isChecked()
    params.add(new BasicNameVal
params.add(new BasicNameValuePa
params.add(new BasicNameValuePa
params.add(new BasicNameValuePa
String resultServer  = getHttp

JSONObject c;
try{
      c = new JSONObject(results
      if(c.getString("Status").e
          Toast.makeText(Registe
 } catch (JSONException e) {
      printStackTrace();
```

Conclusions

Human–machine interfaces in AWS are important for the exercise of human control but there are significant challenges and considerations of design and use before that possibility of control is meaningfully afforded to the operator.

This report has unpacked several aspects of human–machine interfaces, while integrating the discussion in the context of autonomy and human–machine interaction. It used several insights from the autonomous vehicles industry, where considerations of controllability have advanced significantly in the past decade.

General conclusions

- The study of HMIs, as subsystems of AWS, reveals the complexity of human control—as an ability embedded through interface design, cultivated through the process of training, and aided (or compromised) by specific technological properties (e.g., XAI).
- The introduction of autonomous functions and AI, particularly ML, in weapon systems, expands the options and modalities for human–machine interaction; it renders the design and development of human–machine interfaces highly complex, which translates into the need for new kinds of training of human operators.

Interface design

- Since the 1950s, approaches to interface design have evolved from *technology- to user/human- to interaction/human–AI teaming*-centred, reflecting the scaling up of autonomy and autonomous functions.
- It is important for the GGE on LAWS to consider what these paradigms mean from a policy perspective as they go beyond technical upgrades and reflect deeper shifts in human–machine interaction, with direct implications for human control.

Training

- Training of human operators, an important element for human control, comes with new challenges in the context of autonomy in AWS, and as interfaces become more complex.
- Training requires a clearer understanding of AWS limitations, functional allocations, and system failures, and it must address common behavioural factors (e.g., complacency) while maintaining a clear understanding of responsibilities and accountability.

Explainable AI

- While touted as a way of mitigating risks of mistrust in the technology by introducing more understandability and predictability, XAI remains a limited solution for the inherent transparency and explainability problems of autonomous systems. UNIDIR's future research will tackle this topic in greater detail.
- Efforts towards more explainability and transparency in AI-enabled systems must be pursued with due consideration for military demands across classes of operations, contexts, and needs of users. This must be accompanied by thorough research on interface design that best represents the transparency information and the system's brittleness, in order to calibrate expectations and to enhance trust in the technology.

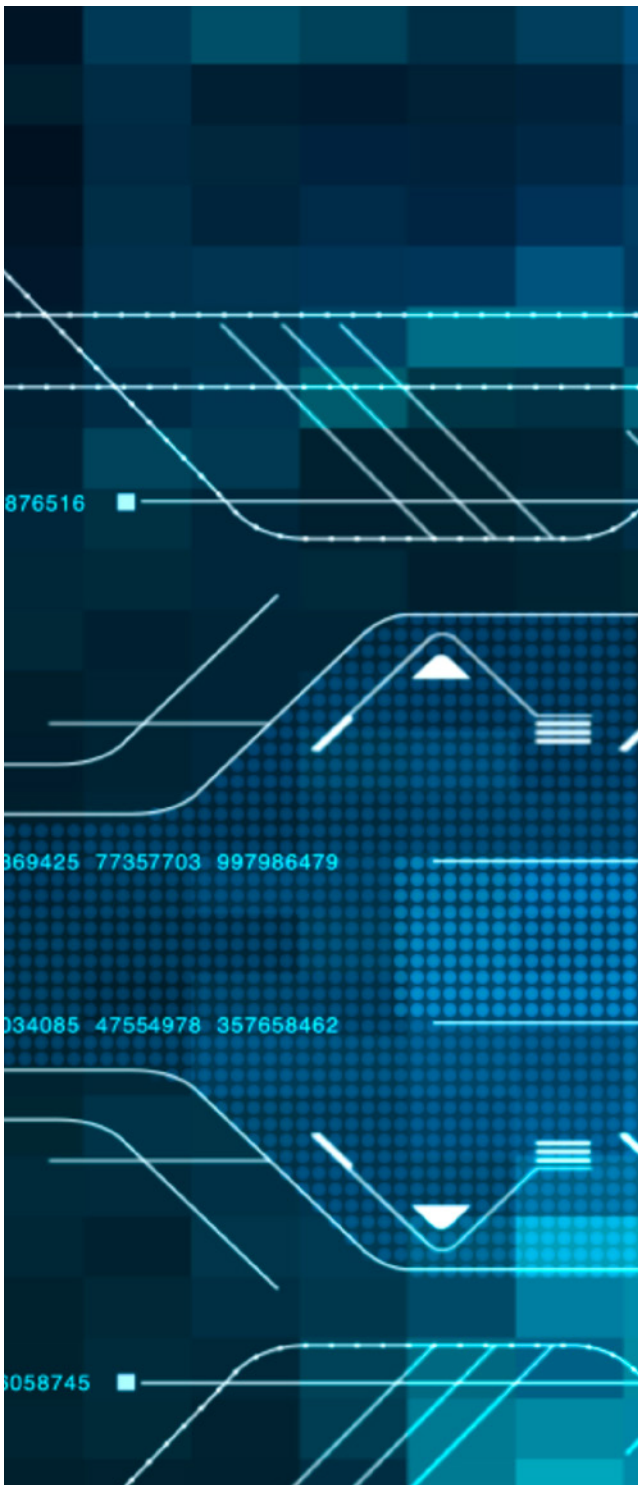
Policy recommendations

This report suggests that the ongoing and future work of the GGE on LAWS should aim to:

Conduct thorough and granular discussions of issues pertaining to human–machine interaction in the context of autonomy, and the role of interfaces in human control. This should include the interconnected aspects of interface design and personnel training in the context of intelligent systems, issues of explainability of AI systems, and trust in the technology.

Articulate more clearly the expectations and objectives related to human control that should guide future development of interfaces for AI systems. The Group's deliberations have been instrumental in framing the issue of human control to the technical community, but more elaboration is needed to ensure that the meaning of human control is more clearly defined, and relevant across types of weapons systems and operational contexts.

Discuss the implications of human–AI teaming for human control over AWS. The metaphor of 'teaming' does not denote an equal status between humans and AI systems, and maintains that humans 'remain in charge'; however, it does prompt new questions about the meaning of human control. Inputs from technical experts are essential in exploring the implications and challenges of this paradigm.



Annex A

Excerpts related to human–machine interaction and human control from GGE on LAWS Reports

Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems CCW/GGE.1/2019/3	**25 September 2019**

21. On the agenda item 5 (c) "Further consideration of the human element in the use of lethal force; aspects of human-machine interaction in the development, deployment and use of emerging technologies in the area of lethal autonomous weapons systems" the Group concluded as follows:

Human responsibility for the use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems can be exercised in various ways across the life-cycle of these weapon systems and through human-machine interaction.

23. On the agenda item 5 (d) "Review of potential military applications of related technologies in the context of the Group's work" the Group concluded as follows:

(…)

(b) Risk mitigation measures can include: rigorous testing and evaluation of systems, legal reviews, readily understandable human-machine interfaces and controls, training personnel, establishing doctrine and procedures, and circumscribing weapons use through appropriate rules of engagement;

Annex IV
Guiding Principles

(c) Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole;

(g) Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems;

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System Chairperson's Summary *Advance Copy* CCW/GGE.1/2020/WP.7	19 April 2021

A. An exploration of the potential challenges posed by emerging technologies in the area of lethal autonomous weapons systems to international humanitarian law

1. Summary of inputs and exchanges

9. Some States have already enacted national legislation to ensure a human is always accountable for decisions on the development and use of weapons. On guiding principle (d), various measures could promote accountability, including rigorous testing and training, establishing procedures and doctrines, and using the weapon system in accordance with established training, doctrine and procedures. [...]

C. Further consideration of the human element in the use of lethal force; aspects of human-machine interaction in the development, deployment and use of emerging techno logies in the area of lethal autonomous weapons systems

1. Summary of inputs and exchanges

27. States should ensure that the use of force must reflect human agency and human intention and that the judgements required to authorize the use of armed force must be made by humans. Context-specific human decisions are necessary to ensure compliance with IHL. Human operators, particularly in the chain of command and control, must have sufficient knowledge and understanding of a system to be confident that it will function as intended in a particular attack.

28. Human-machine interaction has consistently been highlighted by many as a cornerstone on which to build a future operational and normative framework. Many commentaries considered that guiding principle (c) was of primary importance to the work of the group. Several viewed that this principle necessitated further work to determine the type and extent of human involvement required in the use of emerging technologies in the area of LAWS. There may not necessarily be a "one size fits all" set of parameters for human-machine interaction; the requirements for such interaction may instead be dependent on the operational context and the weapon system's characteristics and may need to be determined on a case by case basis. One possible objective of human-machine interaction could be to ensure that humans retain control of the weapons they deploy and operate and the consequences that result. The elaboration of good practices with human-machine interaction that could strengthen compliance with IHL could be valuable. Human-machine interaction may need to be considered at every stage of a weapon system's lifecycle.

29. There was significant discussion of the importance of the concept of "human control". Measures based on a concept of human control could require considerations based on the specific characteristics of a weapon, on the operational environment, on the time-frame of autonomous operation, scope of movement over an area and on human-machine interaction. **Such measures could also specify: the degree of predictability required in a weapon system's behaviour; the required degree of training and understanding of a weapon system; and the ability of a human to deactivate or override the operation of a weapon system. A deactivation requirement, however, may go beyond what States require in currently deployed weapons. Effective human control, involvement or judgment may not necessarily equate to direct, manual control but rather contextual factors including boundaries placed on the weapon and environment of use, and requirements for human-machine interaction**. *[emphasis added]* Further work is needed within the Group to understand various aspects of human control, including the type and extent required for compliance with IHL across all stages of a weapon system's life cycle. The exchange of domestic policies and best practices relevant to this principle could be useful.

30. The ability to constrain a system through setting boundaries on, among other things, its duration of operation, range of operation and the functions that can operate autonomously, and hence determine whether the weapon-system's use could be lawful, was considered as relevant by several delegations. Human operators and commanders need a sufficient understanding of the machines they operate and the algorithms that control the machines' functioning to exercise appropriate judgement and ensure that the use of weapon systems is consistently within applicable international law; hence, control may need to be fully informed to be effective. There also needs to be an understanding of the operational environment. Human control/involvement/judgement might be contingent on the ability to intervene in the operation of a weapon, once activated, though there might also always be a point after which human intervention in a weapon's operation was not possible. Finally, it was noted that human control/involvement/judgement needed to be reasonably temporally proximate to an attack, to remain valid.

2. Possible elements for consensus recommendations

31. Taking into account the guiding principles and the conclusion of the group on agenda item 5 (c) of its report CCW/GGE.1/2019/3, paragraph 21, the group considered the following elements as a possible basis for consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems:

(a) Human responsibility for the use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems can be exercised in various ways across the life-cycle of these weapon systems and through human-machine interaction.

3. Areas for possible future work with a view to arriving at additional elements for consensus recommendations

32. The group could pursue further work on the development of criteria on human control necessary to ensure that the use of emerging technology in the area of LAWS can be limited as required by IHL, including through: (i) operational constraints on the weapon system, (ii) environmental and temporal constraints bounding the operation of the weapon system, and (iii) standards of human-machine interaction, to ensure that all uses of force are meaningfully directed by human operators and commanders.

33. The group could examine methods for assessing the adequacy of constraints and safeguards for ensuring effective human control, involvement or judgment over the employment of weapons based on emerging technologies in the area of LAWS.

D. Review of potential military applications of related technologies in the context of the Group's work

[…]

2. Possible elements for consensus recommendations

37. Taking into account the guiding principles and the conclusions of the group on agenda item 5 (d) of its report CCW/GGE.1/2019/3, paragraphs 23 (a) to (c), the group considered the following elements as a possible basis for consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems:

[…]

b) Risk mitigation measures can include: rigorous testing and evaluation of systems; legal reviews; readily understandable human-machine interfaces and controls; training personnel; establishing doctrine and procedures; and circumscribing weapons use through appropriate rules of engagement.

F. Consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems

[…]

43. In its work, the Group considered elements in relation to the clarification, consideration and development of aspects of the normative and operational framework:

(a) Possible elements that address the normative aspects of the framework can provide clarity regarding how principles and rules of applicable international law, including IHL, apply to emerging technologies in the area of lethal autonomous weapons systems. These could include: the applicability of IHL to States, parties to armed conflict and individuals and their responsibility for adhering to obligations under IHL; the necessity to apply IHL requirements and principles through a chain of command by humans; the necessity for comprehensive, context-based human judgement to ensure compliance with IHL; the applicability of legal reviews of new weapons to emerging technologies in the area of lethal autonomous weapons systems; the specification that it is inherently unlawful to use weapon systems that cannot reliably or predictably perform their functions in accordance with the intention of a human operator and commander to comply with IHL requirements and principles; and that weapon systems based on emerging technologies in the area of LAWS that cannot be used in compliance with IHL should be specifically prohibited;

(b) Possible elements that address the operational aspects of the framework can specify how States should implement principles and rules of IHL with respect to emerging technologies in the area of lethal autonomous weapons systems, as well as how they should cooperate towards this end, could include: ensuring individual responsibility for the employment of weapons systems based on emerging technologies in the area of LAWS; ensuring that a human operator or commander exercises judgement over attacks, including through certain operational constraints on weapon characteristics and environment of use, and requirements for human-machine interaction; […].

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System Geneva, 3–13 August, 24 September–1 October and 2–8 December 2021 Report of the 2021 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems CCW/GGE.1/2021/3	22 February 2022

II. Application of international law

General commitments

26. States should commit to exercise appropriate human involvement throughout the life-cycle of the weapons system that is sufficient to ensure human judgment and control necessary in the circumstances to comply with international humanitarian law over the use of all other types of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems. This may include, but is not limited to:

(a) Limits on the type of target;

(b) Limits on the duration, geographical scope and scale of use;

(c) Requirements for human–machine interaction and necessary intervention or deactivation; or
(d) Clear procedures to ensure that human operators are informed and capable of controlling the weapon systems.

Human-machine interaction

29. The following specific practices in human-machine interaction may contribute to the implementation of international humanitarian law, effective accountability and the mitigation of risks posed by weapon systems based on emerging technologies in the area of lethal autonomous weapon systems:

(a) Human commanders and operators make decisions about the deployment and use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems with information reasonably available at the time to ensure that force will be used in accordance with international law, including information about the potential targets, the capabilities and characteristics of the weapon to be used and the context in which the weapon is deployed.

(b) **Human commanders and operators should be able to properly assess the effects of using a weapons system based on emerging technologies in the area of lethal autonomous weapons systems prior to use.** *[emphasis added]*

(c) **Human commanders and operators and other relevant personnel are trained, to ensure that the weapons systems based on emerging technologies in the area of lethal autonomous weapons systems are deployed and used in conformity with international humanitarian law.** *[emphasis added]*

Risk mitigation

40. Risk mitigation measures to help minimize incidental loss of life, injuries to civilian and damage to civilian objects resulting from the use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems may include, *inter alia*: (a) incorporating self-destruct, self-deactivation, or self-neutralization mechanisms into weapon systems; (b) measures to control the types of targets that the system can engage; (c) measures to control the duration and geographical scope of the weapons system; and((d) **clear procedures for trained human operators to activate or deactivate functions in weapons systems**. *[emphasis added]*

Bibliography

Adams, J. 2002. 'Critical Considerations for Human-Robot Interface Development.' *AAAI Technical Report FS-02-03*. As of 20 June 2022: https://www.aaai.org/Papers/Symposia/Fall/2002/FS-02-03/FS02-03-001.pdf

Ahlstrom, Ulf, Friedman-Berg Ferne J. 2006. 'Using eye movement activity as a correlate of cognitive workload.' *International Journal of Industrial Ergonomics* 36(7): 623–636. https://doi.org/10.1016/j.ergon.2006.04.002

Bahner, Elin J, Anke-Dorothea Hüper & Dietrich Manzey. 2008. 'Misuse of automated decision aids: Complacency, automation bias and the impact of training experience.' *International Journal of Human-Computer Studies* 66: 688-699. doi:10.1016/j.ijhcs.2008.06.001

Bainbridge, Lisanne. 1983. 'Ironies of Automation.' *Automatica* 19 (6):775-779.

Beauxis-Aussalet, Emma, Michael Behrisch, Rita Borgo, Duen Horng Chau, Christopher Collins, David Ebert, Mennatallah El-Assady, Alex Endert, Daniel A. Keim, Jorn Kohlhammer, Daniela Oelke, Jaakko Peltonen, Maria Riveiro, Tobias Schreck, Hendrik Strobelt, Jarke J van Wijk & Theresa-Marie Rhyne. 2021. 'The Role of Interactive Visualization in Fostering Trust in AI.' *IEEE Computer Society*, November/December 2021. DOI: 10.1109/MCG.2021.3107875

Boardman, M., F. Butcher. 2019. 'An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenge of Achieving It.' STO-MP-IST-178 *NATO Report*, 1-16. As of 20 June 2022: https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-178/MP-IST-178-07.pdf

Boulanin, V., N. Davison, N. Goussac, M. Peldán Carlsson. 2020. 'Limits on Autonomy in Weapon Systems. Identifying Practical Limits of Human Control.' *Stockholm International Peace Research Institute (SIPRI) & International Committee of the Red Cross (ICRC)*. As of 20 June 2022: https://www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf

Burt, Chris. 2020. 'Biometrics, touch controls and AI for self-driving training enter growing automotive technology market.' *Biometric Update*. As of 20 June 2022: https://www.biometricupdate.com/202007/biometrics-touch-controls-and-ai-for-self-driving-training-enter-growing-automotive-technology-market

Chandler, Katherine. 2021. 'Does Military AI Have Gender? Understanding Bias and Promoting Ethical Approaches in Military Applications of AI', Geneva: UNIDIR, As of 20 June 2022: https://www.unidir.org/publication/does-military-ai-have-gender-understanding-bias-and-promoting-ethical-approaches

Coeckelbergh, Mark. 2013. 'Drones, information technology, and distance: mapping the moral epistemology of remote fighting.' *Ethics and Information Technology* 15: 87-98. DOI 10.1007/s10676-013-9313-6

Cummings, Mary L. 2006. 'Automation and Accountability in Decision Support System Interface Design.' *The Journal of Technology Studies* 32 (1):23-31. doi.org/10.21061/jots.v32i1.a.4

Davis, Thomas W., Michael Sage Jessee & Anthony W. Morriss. 2017. 'Interface Design: Dynamic Interfaces and Cognitive Resource Management.' In *Designing Soldier Systems. Current Issues in Human Factors*, edited by Pamela Savage-Knepshield, John Martin, John Lockett III & Laurel Allender. Boca Raton: CRC Press.

Development, Concepts and Doctrine Centre (DCDC). UK Ministry of Defence. 2018. 'Joint Concept Note (JCN) 1/18. Human-Machine Teaming.' As of 20 June 2022: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf

Doshi-Velez, F., B. Kim. 2017. 'Towards a Rigorous Science of Interpretable Machine Learning.' *arXiv*. As of 20 June 2022: https://arxiv.org/abs/1702.08608

Endsley, Mica R. 1995. 'Toward a Theory of Situation Awareness in Dynamic Systems.' *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (1): 32-64. https://doi.org/10.1518%2F001872095779049543

Endsley, Mica R., Esin O. Kiris. 1995. 'The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (2): 381-394. https://doi.org/10.1518/001872095779064555

Endsley, Mica R., Betty Bolte & Debra G. Jones. 2003. *Designing for Situation Awareness. An Approach to User-Centered Design*. London: CRC Press.

Endsley, Mica R. 2013. 'Situation Awareness-Oriented Design.' In *The Oxford Handbook of Cognitive Engineering*, edited by John D Lee. & Alex Kirlik. New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780199757183.001.0001

Endsley, Mica R. 2015. 'Situation Awareness Misconceptions and Misunderstandings.' *Journal of Cognitive Engineering and Decision Making* 9 (1): 4-32. DOI: 10.1177/1555343415572631

Endsley, Mica R. 2017. 'From Here to Autonomy: Lessons Learned from Human-Automation Research.' *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59 (1): 5-27. https://doi.org/10.1177/0018720816681350

Feuerriegel, Stefan, Rúben Geraldes, Artur Gonçalves, Ziqi Liu & Helmut Prendinger. 2021. 'Interface design for human-machine collaborations in drone management.' *IEEE Access* 9, 107462-107475. https://doi.org/10.1109/ACCESS.2021.3100712

Flemisch, Frank O, Julian Schindler, Johann Kelsch, Anna Schieben & Daniel Damböck. 2008. 'Some Bridging Methods towards a Balanced Design of Human-Machine Systems, Applied to Highly Automated Vehicles.' *Applied Ergonomics International Conference*. As of 20 June 2022: https://elib.dlr.de/57623

Flemisch, Frank, Eugen Altendorf, Yigiterkut Canpolat, Gina Weßel, Marcel Baltzer, Daniel Lopez, Nicolas Daniel Herzberger, Gudrun Mechthild Irmgard Voß, Maximilian Schwalm & Paul Schutte. 2017. 'Uncanny and Unsafe Valley of Assistance and Automation: First Sketch and Application to Vehicle Automation.' In *Advances in Ergonomic Design of Systems, Products and Processes*, edited by Schlick, Christopher Marc, Sönke Duckwitz, Frank Flemisch, Martin Frenz, Sinem Kuz, Alexander Mertens & Susanne Mütze-Niewöhner. Conference Proceedings. Springer. DOI:10.1007/978-3-662-53305-5_23

Georg, J-M., F. Diermeyer. 2019. 'An Adaptable and Immersive Real Time Interface for Resolving System Limitations of Automated Vehicles with Teleoperation.' 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). As of 20 June 2022: https://ieeexplore.ieee.org/document/8914306/authors#authors

Gillan, Douglas J., Jennifer Riley & Patrick McDermott. 2017. 'The Cognitive Psychology of Human-Robot Interaction.' In *Human-Robot Interactions in Future Military Operations*, edited by Michael Barnes & Florian Jentsch. Boca Raton: CRC Press.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. 2019. 'Explaining Explanations: An Overview of Interpretability of Machine Learning.' *The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018)*. https://arxiv.org/abs/1806.00069

Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS). 2018a. *Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles*. UN Document CCW/GGE.2/2018/WP.1.

------. 2018b. *Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Submitted by France*. UN Document CCW/GGE.2/2018/WP.3.

------. 2018c. *Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Submitted by the United States*. UN Document CCW/GGE.2/2018/WP.4.

------. 2019a. *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. UN Document CCW/GGE.1/2019/3.

------. 2019b. *Final Report of the 13-15 November 2019 Session*. UN Document CCW/MSP/2019/9.

------. 2020. *United Kingdom Expert paper: The human role in autonomous warfare*. UN Document CCW/GGE.1/2020/WP.6.

------. 2021. *Chairperson's Summary*. UN Document CCW/GGE.1/2020/WP.7.

Hartwich, Franziska, Cornelia Hollander, Daniela Johannmeyer, Josef F. Krems, 2021. 'Improving Passenger Experience and Trust in Automated Vehicles Through User-Adaptive HMIs: "The More the Better" Does Not Apply to Everyone.' *Frontiers in Human Dynamics* 3 (Article 669030): 1-17. doi: 10.3389/fhumd.2021.669030

Hawley, J. K., A. L. Mares & C. A. Giammanco. 2005. 'The Human Side of Automation: Lessons for Air Defense Command and Control.' AMR-TR-3468. *Army Research Laboratory*. As of 20 June 2022: https://apps.dtic.mil/sti/pdfs/ADA431964.pdf

Hawley, John K. 2017. 'Patriot Wars. Automation and the Patriot Air and Missile Defense System.' *Center for a New American Security*. As of 20 June 2022: https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf?mtime=20170106135013&focal=none

Hawley, John K., Anna L. Mares. 2017. 'Human Performance Challenges for the Future Force: Lessons from Patriot after the Second Gulf War.' In *Designing Soldier Systems. Current Issues in Human Factors*, edited by Pamela Savage-Knepshield, John Martin, John Lockett III & Laurel Allender,. Boca Raton: CRC Press.

Hoffman, Robert R, Paul J. Feltovich, Stephen M. Fiore, Gary Klein & David Ziebell. 2009. 'Accelerated Learning.' *IEEE Intelligent Systems* 24 (2): 18-22. DOI: 10.1109/MIS.2009.21

Hoffman, Robert R., Paul Ward, Paul J. Feltovich, Lia DiBello, Stephen M. Fiore & Dee H. Andrews. 2014. *Accelerated Expertise. Training for High Proficiency in a Complex World*. New York: Psychology Press.

Hoffman, R. R., S. T. Mueller, G. Klein & J. Litman. 2018. 'Metrics for Explainable AI: Challenges and Prospects.' United States Air Force Research Lab. *arXiv*. arXiv:1812.04608

Holland Michel, A. 2020. 'The Black Box, Unlocked. Predictability and Understandability in Military AI.' Geneva: UNIDIR. As of 20 June 2022: https://unidir.org/publication/black-box-unlocked

Hollnagel, Erik, David D. Woods. 2005. *Joint Cognitive Systems. Foundations of Cognitive Systems Engineering*. Boca Raton: CRC Press.

Hou, Ming, Simon Banbury & Catherine Burns. 2015. *Intelligent Adaptive Systems. An Interaction-Centered Design Perspective*. Boca Raton: CRC Press.

Hou, M. 2020. 'IMPACT: A Trust Model for Human-Agent Teaming.' *2020 IEEE International Conference on Human-Machine Systems (ICHMS).* As of 24 June 2022: https://ieeexplore.ieee.org/document/9209519

Hou, Ming, Yingxu Wang, Ljiljana Trajkovic, Konstantinos N. Plataniotis, Sam Kwong, MengChu Zhou, Edward Tunstel, Imre J. Rudas, Janusz Kacprzyk & Henry Leung. 2022. 'Frontiers of Brain-Inspired Autonomous Systems: How Does Defense R&D Drive the Innovations?' *IEEE Systems, Man & Cybernetics Magazine*. DOI: 10.1109/MSMC.2021.3136983

Hulstaert, Lars. 2018. 'Understanding Model Predictions with LIME.' *Towards Data Science*, 11 July 2018. As of 20 June 2022: https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b

International Organization for Standardization (ISO). 2013. ISO/TS 20282-2:2013. As of 20 June 2022: https://www.iso.org/standard/62733.html

International Organization for Standardization (ISO). 2019. ISO 9241-220:2019. As of 20 June 2022: https://www.iso.org/standard/63462.html

International Panel on the Regulation of Autonomous Weapons (iPRAW). 2019. *Focus* on *Human Control*. No. 5, August 2019. As of 20 June 2022: https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf

Janssen, Christian P., Stella F. Donker, Duncan P. Brumby & Andrew L. Kun. 2019. 'History and future of human-automation interaction.' *International Journal of Human-Computer Studies* 131: 99-107. https://doi.org/10.1016/j.ijhcs.2019.05.006

Johnson, Matthew, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, Birna van Riemsdijk & Maarten Sierhuis. 2011. 'The Fundamental Principle of Coactive Design: Interdependence must Shape Autonomy.' In *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, edited by Marina de Vos, Nicoletta Fornara, Jeremy V. Pitt & George Vouros. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-21268-0_10

Johnson, Matthew, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk & Maarten Sierhuis. 2014. 'Coactive Design: Designing Support for Interdependence in Joint Activity.' *Journal of Human-Robot Interaction* 3 (1): 43-69. DOI:10.5898/JHRI.3.1.Johnson

Johnson, Matthew, Jeffrey M. Bradshaw & Paul J. Feltovich. 2018. 'Tomorrow's Human-Machine Design Tools: From Levels of Automation to Interdependencies.' *Journal of Cognitive Engineering and Decision Making* 12 (1): 77-82. doi.org/10.1177/1555343417736462

Kunze, Alexander, Stephen J. Summerskill, Russell Marshall & Ashleigh J. Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces.' *Ergonomics* 62 (3): 345-360. https://doi.org/10.1080/00140139.2018.1547842

Linardatos, Pantelis, Vasilis Papastefanopoulos & Sotiris Kotsiantis. 2021. 'Explainable AI: A Review of Machine Learning Interpretability Methods.' *Entropy* 23 (18): 1-45. https://dx.doi.org/10.3390/e23010018

Lee, John D., Katrina A. See. (2004). 'Trust in Automation: Designing for Appropriate Reliance.' *Human Factors* 46 (1): 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

McNeese, Nathan J, Beau G. Schelble, Lorenzo Barberis Canonico & Mustafa Demir. 2021. 'Who/What is My Teammate? Team Composition. Considerations in Human-AI Teaming.' *IEEE Transactions on Human-Machine Systems*, 51(4). https://doi.org/10.48550/arXiv.2105.11000

Metzger, Ulla, Raja Parasuraman. 2001. 'The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control versus Passive Monitoring.' *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43 (4): 519-528. https://doi.org/10.1518/001872001775870421

Mueller, S. T., R. R. Hoffman, W. Clancey, A. Emrey & G. Klein. 2019. 'Explanations in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI.' DARPA XAI Program. *arXiv*. arXiv:1902.01876

Mueller, S. T., E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. Mamun & W. J. Clancey. 2020. 'Principles of Explanation in Human-AI Systems.' *Association for the Advancement of Artificial Intelligence*. arXiv:2102.04972

National Academies of Sciences, Engineering and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press. doi.org/10.17226/26355

Onnasch, Linda, Christopher D.Wickens, Huiyang Li, Dietrich Manzey. 2014. 'Human Perfomance Conseqsuences of Stages and Levels of Automation: An Integrated Meta-Analysis.' *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (3):476-488.

Oury, Jacob D., Frank E. Ritter. 2021. *Building Better Interfaces for Remote Autonomous Systems. An Introduction for Systems Engineers*. eBook: Springer. https://doi.org/10.1007/978-3-030-47775-2

Parasuraman, Raja, Victor Riley. 1997. 'Humans and Automation: Use, Misuse, Disuse, Abuse.' *Human Factors* 39 (2):230-253. doi.org/10.1518/001872097778543886

Parasuraman, Raja. 2000. 'Designing automation for human use: empirical studies and quantitative models.' *Ergonomics* 43 (7):931-951. https://doi.org/10.1080/001401300409125

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan & Hanna Wallach. 2021. 'Manipulating and Measuring Model Interpretability.' *CHI Conference on Human Factors*, May 8–13, 2021, Yokohama, Japan. https://doi.org/10.1145/3411764.3445315

Riley, Jennifer M., Laura D. Strater, Sheryl L. Chappell, Erik S. Connors, Mica Endsley. 2017. 'Situation Awareness in Human-Robot Interaction: Challenges and User Interface Requirements.' In. *Human-Robot Interactions in Future Military Operations*, edited by Michael Barnes & Florian Jentsch. Boca Raton: CRC Press.

Russell, Stuart, Peter Norvig. 2022. *Artificial Intelligence: A Modern Approach (4th edition)*, Harlow: Pearson Education.

Santoni de Sio, F., J. van den Hoven. 2018. 'Meaningful Human Control over Autonomous Systems: a Philosophical Account.' *Frontiers in Robotics and AI* 5(15), 1-14. doi.org/10.3389/frobt.2018.00015

Sartori, Laura, Andreas Theodorou. 2022. 'A Sociotechnical perspective for the future of AI: narratives, inequalities, and human control.' *Ethics and Information Technology* 24 (4): 1-11. https://doi.org/10.1007/s10676-022-09624-3

Savage-Knepshield, Pamela A. 2017. 'Soldier-centered Design and Evaluation Techniques.' In *Designing Soldier Systems. Current Issues in Human Factors*, edited by Pamela Savage-Knepshield, John Martin, John Lockett III & Laurel Allender. Boca Raton: CRC Press.

Scharre, Paul. 2018. *Army of None: Autonomous Weapons and the Future of War*. New York: W.W. Norton & Company.

Schwarz, Elke. 2021. 'Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control.' *The Philosophical Journal of Conflict and Violence* 5 (1): 53-72. DOI:10.22618/TP.PJCV.20215.1.139004

Schultze, Ulrike, Wanda J. Orlikowski. 2010. 'Research Commentary: Virtual Worlds: A Performative Perspective on Globally Distributed, Immersive Work.' *Information Systems Research* 21 (4):810-821. doi 10.1287/isre.1100.0321

Siebert, Luciano Cavalcante, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van den Hoven, Deborah Forster & Reginald L. Lagendijk. 2022. 'Meaningful human control: actionable properties for AI system development.' *AI and Ethics* (18 May 2022). https://doi.org/10.1007/s43681-022-00167-3

Sperrle, Fabian, Astrik Jeitler, Jürgen Bernard, Daniel A Keim & Mennatallah El-Assady. 2020. 'Learning and Teaching in Co-Adaptive Guidance for Mixed-Initiative Visual Analytics.' *EuroVis Workshop on Visual Analytics*. Eurographics Proceedings. DOI: 10.2312/eurova.20201088

Spinner, Thilo, Udo Schlegel, Hanna Schafer& Mennatallah El-Assady. 2019. 'explAIner: A Visual Analysis Framework for Interactive and Explainable Machine Learning.' *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.48550/arXiv.1908.00087

Swartz, Luke. 2001. 'Overwhelmed by Technology: How did user interface failures on board the USS *Vincennes* lead to 290 dead?' As of 20 June 2022: http://xenon.stanford.edu/~lswartz/vincennes.pdf

United Nations Institute for Disarmament Research (UNIDIR). 2014. 'The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward.' UNIDIR Resources, 2. As of 15 June 2022: https://unidir.org/publication/weaponization-increasingly-autonomous-technologies-considering-how-meaningful-human

United States Air Force (USAF). 2015. 'System Autonomy in the Air Force – A Path to the Future. Volume I: Human-Autonomy Teaming.' *Office of the Chief Scientist AF/ST TR 15-01*. As of 20 June 2022: https://www.af.mil/Portals/1/documents/SECAF/AutonomousHorizons.pdf?timestamp=1435068339702

United States Department of Defense (US DoD). 2012. 'Autonomy in Weapon Systems.' *Directive 3000.09*, 21 November 2012, Incorporating Change 1, 8 May 2017. As of 20 June 2022: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

United States Mission to International Organisations in Geneva. 2021. Statement at the GGE on LAWS – Agenda item 5(c), 4 August 2021. As of 20 June 2022: https://geneva.usmission.gov/2021/08/04/u-s-statement-at-the-gge-on-laws-agenda-item-5c

United States Department of Defense. 2022. 'Human Systems Integration Guidebook.' Washington DC: Office of the Under Secretary of Defense for Research and Engineering, May 2022. As of 20 June 2022: https://ac.cto.mil/wp-content/uploads/2022/06/HSI_Guidebook_May2022-Cleared.pdf

van der Waa, Jasper, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, & Ioana Cocu. 2021. 'Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations.' *Frontiers in Robotics and AI* 8, Article 640647. doi: 10.3389/frobt.2021.640647

van der Waa, Jasper, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx & Catholijn Jonker. 2020. 'Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach.' *Lecture Notes in Computer Science* Volume 12187. Springer. https://doi.org/10.1007/978-3-030-49183-3_16

Zhang, Rui, Nathan J. McNeese, Guo Freeman & Geoff Musick. 2020. '"An Ideal Human": Expectations of AI Teammates in Human-AI Teaming'. *Proceedings of the ACM on Human-Computer Interaction* 4 (Issue CSCW3), Article 246: 1-25. https://doi.org/10.1145/3432945

HUMAN–MACHINE INTERFACES IN AUTONOMOUS WEAPON SYSTEMS

Considerations for Human Control

IOANA PUSCAS

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

@unidirgeneva @UNIDIR un_disarmresearch