THE BLACK BOX, UNLOCKED

PREDICTABILITY AND UNDERSTANDABILITY IN MILITARY AI

ARTHUR HOLLAND MICHEL

40 UNIDIR

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all the Institute's activities. This study was produced by the Security and Technology Programme, which is funded by the Governments of Germany, the Netherlands, Norway and Switzerland, and by Microsoft. The author wishes to thank the study's external reviewers, Dr. Pascale Fung, Lindsey R. Sheppard, Dr. Martin Hagström, and Maaike Verbruggen, as well as the numerous subject matter experts who provided valuable input over the course of the research process.

Design and layout by Eric M. Schulz.

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 the publication are the sole responsibility of the individual authors. They do not necessary reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

Citation

Holland Michel, Arthur. 2020. 'The Black Box, Unlocked: Predictability and Understandability in Military AI.' Geneva, Switzerland: United Nations Institute for Disarmament Research. doi: 10.37559/SecTec/20/AI1

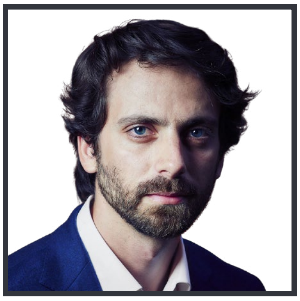
About UNIDIR

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.

IN COMPUTING, *a "black box" is a system for which we know the inputs and outputs but can't see the process by which it turns the former into the latter. Somewhat confusingly, airplane flight recorders are also referred to as black boxes; when an artificially intelligent system is indecipherable, it is like an airplane black box for which we have no key.*

TABLE OF CONTENTS

ABOUT THE AUTHOR



ARTHUR HOLLAND MICHEL is an Associate Researcher with the Security and Technology Programme at UNIDIR. He is the founder of the Center for the Study of the Drone at Bard College—where he was Co-Director from 2012 to 2020—and currently serves as a Senior Fellow focusing on autonomy and advanced surveillance technology at the Carnegie Council for Ethics in International Affairs. He has written widely for popular and academic media about unpiloted systems, artificial intelligence and other emerging security and surveillance technologies. His first book, *Eyes in the Sky: The Secret Rise of Gorgon Stare and How It Will Watch Us All*, was published by Houghton Mifflin Harcourt in 2019. Follow Arthur on Twitter @WriteArthur.

ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
IHL	International Humanitarian Law
LAWS	Lethal Autonomous Weapon Systems
XAI	Explainable Artificial Intelligence

INTRODUCTION

"Predictability" and "understandability" are widely held to be vital qualities of artificially intelligent systems.[1] Put simply: such systems should do what they are expected to do, and they must do so for intelligible reasons. This view represents an important point of common ground among the many different parties to the debate on emerging technologies in the area of lethal autonomous weapon systems (LAWS) and other forms of military AI. Just as the unrestricted employment of a completely unpredictable lethal autonomous weapon system that behaves in entirely unintelligible ways would likely be universally regarded as injudicious and illegal, the use of a perfectly predictable and understandable autonomous weapon system—if such a system existed—probably would not pose many of the central regulatory concerns that underlie the ongoing debate.

This suggests that any path that is ultimately taken to address the use of LAWS and other forms of AI in military applications must account for what is sometimes known as the "black box dilemma" of AI. Indeed, adherence to existing international humanitarian law (IHL), let alone hypothetical new laws, may even hinge on specific measures to ensure that LAWS and other military AI systems do what they are expected to do, and do so for understandable reasons.[2] And yet for the most part in the discourse on LAWS and military AI, predictability and understandability have not yet been treated with the kind of detailed foregrounding that befits an issue of such importance and complexity. This has led to confusion around the technical fundamentals of AI predictability and understandability, how and why they matter, and the potential avenues by which the black box dilemma might be addressed.

This report seeks to resolve these ambiguities by providing a common baseline of knowledge on this subject. Sections 1 and 2 explain what exactly it means to say that an intelligent system is "predictable" and "understandable" (or, conversely, "unpredictable" and "unintelligible") and illustrate that there are various types of understandability and predictability that differ in important ways. Section 3 describes the specific practical reasons why predictability and understandability will be necessary characteristics of LAWS and other military AI at every stage of their development, deployment and post-use assessment. Section 4 lists the factors that will determine the appropriate required level and type of predictability and understandability at each of these stages. Section 5 discusses measures that may be necessary to achieve and assure these levels of predictability and understandability—including training, testing, standards and Explainable AI (XAI) technology. The conclusion proposes five avenues for further inquiry and action for policy stakeholders, militaries and the technical community.

NOTE

While "explainability" is more commonly used in the debate on LAWS, this report opts for the term "understandability". In the scientific literature, "explainability" generally refers only to technical measures that "explain" black box AI systems (see page 21) and not to systems that are inherently interpretable. Furthermore, the word "explainability" unhelpfully implies a degree of human-like agency on the part of the AI system, in the sense that it can "explain" itself like a person. The broader and more neutral term "understandability" covers the technical explainability and interpretability of the AI system while also accounting for the human subject's capacity for understanding—and it does not imply agency on the part of the machine.

```
   > .sf-sub-indicator [
ent .cart-menu .cart-icon-wr
er-outer.transparent header#to
nav .sf-menu > li.current_page_
nav .sf-menu > li.current-menu-
nav > ul > li > a:hover > .sf-su
nav ul #search-btn a:hover span,
nav .sf-menu > li.current-menu-it
hover .icon-salient-cart,.ascend
:1!important;color:#ffffff!impo
arent header#top nav>ul>li.butt
-out-widget-area-toggle a i.
    er-outer.transparent
```

KEY TAKEAWAYS OF THIS REPORT

- There are **three distinct senses of AI un/predictability**: the degree to which a system's technical performance is or is not consistent with past performance, the degree to which any AI or autonomous system's[3] specific actions can (and cannot) be anticipated, and the degree to which the effects of employing an AI system can be anticipated. **(Page 5)**

- Predictability is a function of a system's **technical characteristics**, the kind of **environments and adversaries** to which the system is subjected, and **the degree to which it is understood by its users**. **(Page 7)**

- Understandability is based on a system's **intrinsic interpretability** as well as the human subject's **capacity for understanding**. There are **multiple ways in which an intelligent system can be "understood"**, not all of which are grounded in technical aspects of the system or the human's technical literacy. **(Page 10)**

- **Predictability is not an absolute substitute for understandability, or vice versa**. A combination of both high predictability and high understandability may be the only optimal condition for safe, prudent and compliant use of complex intelligent or autonomous military systems. **(Page 11)**

- Predictability and understandability **are necessary qualities in autonomous weapons and other forms of military AI** for a wide range of reasons **throughout their development, use and assessment**. **(Page 13)**

- **The appropriate level and type of predictability and understandability in these systems will vary widely** according to a range of factors, including the type and criticality of the mission, the kind of environment or input data, and the type of stakeholder assessing or operating the system. **(Page 17)**

- Potential approaches to achieve and assure appropriate predictability and understandability in military AI systems will likely implicate efforts related to **training, testing and standards**. Technical research efforts to build XAI also offer some promise, but this remains a nascent field. **(Page 19)**

1. WHAT IS PREDICTABILITY?

Predictability is the extent to which a system's outputs or effects can be anticipated. In other words, it is the degree to which one can answer the question, *What will this system do?* Within this broad definition, there are three specific senses of predictability—or lack of predictability: "unpredictability"—that diverge in important ways.

In the **technical sense**, "predictability" generally refers to a system's ability to execute a task with the same performance that it exhibited in testing, in previous applications or (in the case of machine learning systems) on its training data. This is a function of how often the system's outputs are correct (accuracy),[4] the extent to which it can achieve the same accuracy over time (replicability, reproducibility), and the extent to which the system can "generalize" to accommodate input data that diverge from the data for which the system was designed or on which it was trained or tested.[5]

AI predictability is not the same as reliability, which refers to the extent to which a system does or does not fail. Even exceptionally reliable systems that fail rarely might still occasionally fail in very unpredictable ways[6] because the range of failures that the system can exhibit is wide. As a result, AI failures can be very difficult to model in advance—a fact that can be compounded by their opacity (see Section 2).[7]

In the **operational sense**, "predictability" refers to the degree to which an autonomous system's individual actions can be anticipated.[8] All autonomous systems exhibit a degree of inherent operational unpredictability, even if they do not fail or the outcomes of their individual action can be reasonably anticipated.[9] This is because, by design, such systems will navigate situations[10] that the operators cannot anticipate.[11] Consider a fully autonomous drone that maps the interior of a network of tunnels. Even if the drone exhibits a high degree of technical predictability and exceptional reliability, those deploying the drone cannot possibly anticipate exactly what it will encounter inside the tunnels, and therefore they will not know in advance what exact actions the drone will take.[12] While technical predictability is solely a function of a system's performance, operational predictability is just as much a function of the characteristics of the environment and mission for which the system is deployed.

Operational unpredictability is particularly inherent in systems designed to handle a wide range of inputs, complex environments and dynamic conditions. Not only is it hard to anticipate what such a system will encounter, it may be difficult (especially in the case of learning-based systems) to anticipate exactly how the system will respond to this environment, because such AI systems may achieve their goals in ways that are not necessarily logical or reasonable by human standards (see Figure 1). This notion of inherent operational unpredictability is elemental to some countries' definition of lethal autonomous weapons and underpins some groups' objections to the development and use of such weapons.[13]

FIGURE 1

Sometimes a high degree of unpredictability is not exactly a bug, but a feature. When Google DeepMind's AI systems famously trounced two champion Go players in 2016 and 2017, they did so in part with moves that experts described as "alien" and "from an alternate dimension".[14] Roman Yampolskiy describes this unavoidable unpredictability of complex AI with the following simple proof: "Suppose that unpredictability is wrong and it is possible for a person to accurately predict decisions of superintelligence. That means they can make the same decisions as the superintelligence, which makes them as smart as superintelligence but that is a contradiction as superintelligence is defined as a system smarter than any person is. That means that our initial assumption was false and unpredictability is not wrong."[15]

The interaction of technical and operational un/predictability gives rise to a **third**, general, meaning of un/predictability: the degree to which the **outcomes or effects of a system's use can be anticipated**. A broad range of overlapping factors determine technical and operational predictability—and thus the predictability of effects—in any given instance of employment. These factors include:

Type of system—There are many types of algorithmic system, and different systems may be more or less predictable in terms of performance, may lend themselves to more or less predictable types of operations,[16] may be more or less likely to engage in unwanted or surprising behaviours,[17] and as a result may fail in more or less predictable ways. The level of computing power available for a system may also affect predictability.[18]

Type of task or function—Broadly speaking, the greater the variety of possible outputs or actions that a system can generate (sometimes known as the system's "decision space"[19]), the harder it is to predict individual outputs or actions. For example, a system that simply detects objects with a yes/no alert may be more predictable than a system that characterizes objects by type (say, a system that distinguishes between airplanes, cars, tanks, people and trees).

System development or training data—All AI systems exhibit a degree of "brittleness": the tendency to fail, sometimes unpredictably, in response to inputs for which they have not been designed or trained.[20] Systems may exhibit brittleness as a result of encountering edge case or corner case conditions at the outer boundaries of the parameters they were designed for[21] or encountering individual out-of-distribution inputs that fall outside the scope of the training data,[22] or when there are broad differences between the training data and the input data: a phenomenon known as "data shift".[23]

Testing—The more extensively a system has been tested against a greater variety of environments or inputs, the less likely the system is, when fielded, to encounter inputs for which its response has never been observed or validated.[24]

Complexity of the environment—The more complex the operating environment to which a system is deployed, the more likely it is that the system will encounter inputs for which it was not specifically trained or tested[25] or will display new behaviours that have not been previously observed or validated (sometimes known as "emergent behaviours"[26]).

Capacity for self-learning—If a machine learning system can adjust itself in real time while executing a mission—a technique that is gaining favour as a means of continuously improving the system's performance and further enabling autonomous operations in complex environments—that system's specific outputs may be harder to predict as it may acquire new unanticipated behaviours that have not been tested.[27]

Scale and length of the deployment—As the scale of the environment increases and the length of the operation grows, the more likely a system is to encounter inputs for which it was not specifically trained or tested,[28] exhibit certain behaviours that cannot be individually anticipated, or (in the case of self-learning systems) acquire new unverified capabilities. For example, a fixed air-defence system that only targets a small slice of airspace for a limited window of time is far less likely to encounter edge cases or run through a long series of unpredictable actions[29] than an autonomous drone that can rove across thousands of square miles in search of targets. This is why some groups have suggested that implementing strict constraints on where an autonomous weapon operates, as well as how long it operates, could help counteract unpredictability in such systems.[30]

Data quality—Low-quality or insufficient input data may give rise to system failures or other outputs that could increase unpredictability.[31] This factor is particularly important, as adversarial actors may intentionally present autonomous systems with low-quality, distorted, falsified, spoofed, or other out-of-distribution inputs to confuse and defeat these systems.[32]

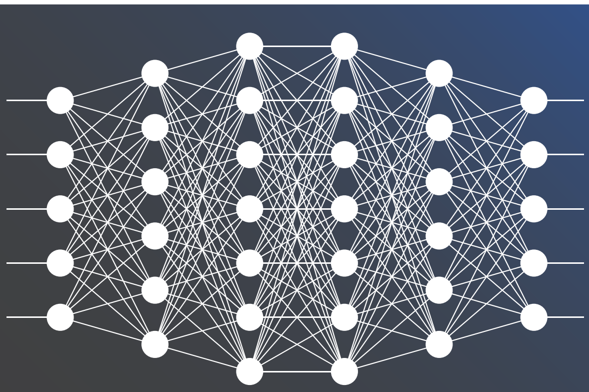
Number of interacting systems—When multiple AI systems work together or in tandem (as in a swarm) or interact with other complex or intelligent systems in the environment, the effects of these interactions may be exponentially more difficult to predict.[33]

2. WHAT IS UNDERSTANDABILITY?

Understandability refers to the degree to which any given system can be understood by any given person. Whereas a system's predictability relates to the question *What will the system do?* understandability relates to the question *Why does it do it?*[34]

Some forms of AI can be indecipherable even to the humans who built them, let alone the non-technical individuals who will use them or be subject to their actions.[35] The issue of understandability has received much attention in recent years because of the growing use of these kinds of systems in critical applications where a lack of understanding of why an AI system is doing what it is doing could pose grave risks. Without an understanding of why a system is doing what it does, it may be difficult to assess if it is operating correctly or if it is failing, and it may be difficult to anticipate the system's future actions—meaning that better understanding improves predictability. There are a variety of ways a system can be understood, and different forms of understanding may be more or less appropriate for certain contexts.[36]

Although understandability is particularly challenging in relation to learning-based systems (see insert), it is an important consideration for all types of AI. Even simpler rules-based AI systems might be dangerously incomprehensible to a layperson or a user who has never interacted with the system before. Other terms sometimes used in reference to the concept of understandability include "transparency", "interpretability" and "intelligibility".[37]



A diagram of a neural network architecture

The individual outputs of machine learning systems may be especially difficult to anticipate and understand. For example, a fairly standard neural network-based computer vision system that can identify animals does so using a probabilistic process that assigns millions of weights to the features detected in every given image. This is a level of complexity that no human mind can fully grasp. Nor is there necessarily anything broadly intuitive about the process.[38] By design, machine learning systems make their own "rules" for how to achieve the goals that are set for them.[39]

As a result, it is hard to tell what features, or "artifacts", in the input data the system is likely to draw on to reach a conclusion.[40] Such a vision system might, for instance, learn to distinguish between huskies and wolves not on the basis of the shape of the animals themselves, but rather on the basis of the image background: if most of the wolf images that the system is trained on have a snowy background, the system might learn to "identify" wolves by simply detecting the presence of white in any given image.[41] This may give developers the impression that the system is highly accurate, but once deployed the system might classify any white image as a wolf and fail to detect wolves that aren't surrounded by snow, making the system more unpredictable. A related phenomenon is the tendency of certain types of learning algorithm to "cheat" to achieve a specified objective—this is known as "specification gaming" or "reward hacking", and may be equally problematic.[42]

2.1 THE TWO FACTORS THAT GO INTO UNDERSTANDABILITY

Broadly speaking, two factors determine understandability: **the features of the AI system** and **the human's capacity for understanding**. Understandability is always a function of the two factors in combination.[43]

2.1.1 Features of the AI System

AI systems vary widely in terms of how intrinsically easy they are to comprehend. For instance, a simple rule-based algorithm with a limited number of elements and a clear intuitive logic is likely to be more understandable than a neural network with hundreds of hidden layers. Systems that are intrinsically understandable are sometimes referred to as "interpretable" or "transparent" models. Systems that are not inherently understandable are sometimes referred to as "black box" or "opaque" AI.[44]

In some cases, extrinsic explanation tools may be added to an opaque AI model to make it more understandable. Such tools either provide specific local explanations for how the system produced a specific given output (these are sometimes known as "post hoc" explanations[45]) or a global explanation for how the AI system produces its outputs in general.[46] These tools are the subject of a growing body of technical research (see page 21).

2.1.2 Human Capacity for Understanding

Every human who is assessing or operating an AI system has a different unique capacity for understanding. This is based on their level of technical expertise, their knowledge of that system's past performance, their knowledge of the system's training data,[47] their understanding of the environment to which the AI system is being deployed and the data it will ingest, and the level of attention they can give the system in an operation (their "cognitive load").[48]

A human's capacity for understanding is not always grounded in technical literacy or direct insight into a system's architecture. If an opaque system is sufficiently predictable in its performance, and if the human has spent enough time observing the system, this may be sufficient for the human to build a reliable mental model of how the system works, even if they do not understand its neural architecture.[49] (By analogy, most people probably do not know exactly how their toaster works, but they do have a robust mental model of how it turns bread into toast.) While mental models can be useful, they may also have limitations if not accompanied by some level of technical understanding. For example, if a human-machine team encounters conditions that differ significantly from those they have previously encountered, the operator's mental model may not be able to anticipate how exactly the system will respond.[50]

2.2 THE PERFORMANCE-UNDERSTANDABILITY TRADE-OFF

Higher performing AI systems tend to be more complex than less advanced AI systems.[51] For example, simple symbolic systems are generally quite understandable, but they are often too brittle for employment in complex environments. By contrast, a sophisticated learning-based system may be better suited to such environments but it will probably also be less understandable[52] and thus less predictable.[53]

Similarly, the more types of inputs and parameters a system can account for in its processing, the sharper its accuracy may be. But it is also likely to be less understandable than systems that calculate just a handful of parameters across a single data stream.[54] Likewise, if multiple transparent models are fused, their outputs could be unintelligible[55] or, if multiple individual AI agents (for instance, a swarm) operate as a collective network to achieve a goal, the process by which that goal is achieved may be highly uninterpretable.

This performance-understandability trade-off poses a central paradox of AI.[56] As groups seek to employ AI for increasingly critical roles, such as transportation, medicine and warfare, they want systems that have the best possible performance. But those systems may prove challenging for human operators and stakeholders to understand—which in turn poses serious risks.

2.3 HOW UNDERSTANDABILITY AND PREDICTABILITY RELATE

Understandability can improve the human's capacity to anticipate what a system will do, thus improving predictability. Predictability can make it easier for a human to develop a robust mental model of the system, thus improving their understanding. However, given that predictability relates to what a system will do and understandability relates to *why it does it*, **predictability is not an absolute substitute for understandability, or vice versa**. For example, if an exceptionally predictable autonomous system operates in a complex or adversarial environment or processes large volumes of diverse and cluttered data, it is likely to be impossible to know how the system would react to all the unique conditions and inputs it could encounter in any given instance—even if it has performed consistently well in previous cases—without some degree of understandability.[57] Nor would it necessarily be possible to assess, by way of understanding alone, why a given system produced certain outputs or achieved certain effects, or determine if it is likely to behave the same way in future, if that system has not exhibited an appropriate level of predictability in the past. Highly unpredictable systems cannot be accounted for by technical understanding alone.

Therefore, a combination of both high predictability and high understandability may be the only optimal condition for the safe and prudent use of complex AI systems grounded in human responsibility.

2.4 DRAWING FROM THE CIVILIAN AI REALM

In recent years, much of the most advanced discourse and research on understandability and predictability has focused on civilian applications of AI, particularly in areas such as transportation, finance, medicine and security.[58] In these domains, just as in the military domain, there are many reasons to ensure that AI systems can be both anticipated and understood. Much of the technical and analytical work that has already been done in the civilian realm could therefore have direct relevance and even practical utility for those seeking to address these considerations in the military domain.

However, the civilian realm and the military realm also differ in crucial ways. The two domains are subject to different rules. The actors in each domain are motivated by different goals. Mistakes in each sphere have different repercussions, on different scales of magnitude; the algorithms employed in automated command and control software may someday resemble the algorithms used for ride-hailing apps, but the repercussions of even a minor failure in the former are likely to be far more profound than the effects of an error in the latter.

By and large, the environments to which advanced military AI systems are subjected could be much more dynamic and complex than anything faced by equivalent commercial AI systems. In both the physical and digital domains, active battlefields are complex and ever-changing in ways that peacetime civilian environments usually are not. And, crucially, the military environment is far more adversarial than the civilian environment. Many AI and autonomous systems used in conflict will have to contend with active efforts of denial, deception and subterfuge on the part of the adversary.[59]

These differences illustrate that in certain regards, predictability and understandability have a specific and unique relevance to military AI that cannot be fully illuminated or addressed through research and measures geared for civilian AI. Broadly speaking, these unique characteristics of the military domain may make predictability and understandability both more critical and harder to achieve.

3. THE ROLE OF PREDICTABILITY AND UNDERSTANDABILITY

The international community has not yet decided if or how to create and implement rules for AI systems in warfare. But adherence to any of the options currently under consideration—including the operationalization of international guiding principles,[60] new fit-for-purpose rules, an outright ban on LAWS, or simply the existing requirement for adherence to IHL—will hinge on the ability to have systems that are both appropriately predictable and understandable. This is likely to be just as true for the employment of "human-out-of-the-loop" systems as it will for "human-in-the-loop" or "human-on-the-loop" teaming arrangements.[61]

For example, the guiding principles adopted by the Group of Governmental Experts on LAWS in 2019 assert that IHL always applies to the employment of autonomous weapons and that humans are always ultimately responsible for the effects of such employment.[62] This implies, by default, a requirement of understandability and predictability among users of such systems. If a human operator is setting constraints on where and how a lethal autonomous weapon system operates, they would want to have reasonable confidence that the system will comply with those constraints,[63] understand how it will do so,[64] and know how it will respond to confounding or compromised inputs.[65]

By the same token, a legally mandated standard of "meaningful human control"[66] over autonomous systems would imply that the human interacting with the system at any stage of development or employment would have an adequate understanding of how the system works, why it produces given outputs, and what it is likely to do next.[67] For instance, if an operator directing a human-in-the-loop autonomous system approves the system's targeting selections without understanding why it made those selections or how likely it is to strike them accurately, this likely would not count as meaningful or sufficient human control according to most definitions of those terms.[68]

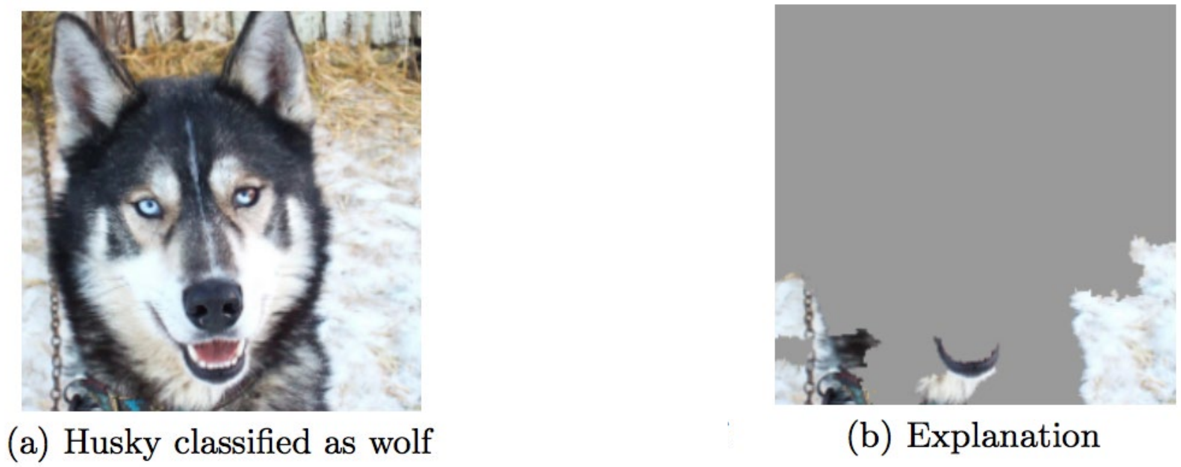
Regardless of which policy options are ultimately applied, assured understandability and predictability are likely to be necessary throughout the entire timeline of the development and employment of autonomous weapon systems[69] and other forms of AI related to combat functions.[70] This section describes exactly why and how predictability and understandability are likely to play important roles at each of these stages.

3.1 BEFORE EMPLOYMENT

Long before a weapon system is fielded, it goes through a variety of processes intended to certify that it will perform in conflict as desired and that it will not have any unintended harmful consequences as a result of flaws, bugs or other issues that could result in failures. Appropriate understandability and predictability will be fundamental to these processes.

In **testing and evaluation**, system understandability could be crucial for debugging and troubleshooting;[71] gauging whether the model is well fitted to its proposed operational environment, flagging issues such as bias[72] or irrelevant artifacts[73] in the training data[74] (see Figure 2); identifying whether the system is achieving its results legitimately rather than by way of irrelevant artifacts or "reward hacking";[75] and highlighting other vulnerabilities[76] that may not have been identified in development. Meanwhile, systems with higher predictability will engender a more precise assessment of whether they will exhibit the same behaviours in the field that they exhibited in testing; in contrast, if

FIGURE 2



(a) Husky classified as wolf

(b) Explanation

Explainability tools that improve understanding could potentially serve to highlight problems with AI systems that might not otherwise be immediately visible. In one well-known experiment, researchers showed how an explainability tool (right) could highlight how a computer vision system that distinguishes huskies from wolves was erroneously identifying wolves based solely on the presence of snow in the images (since most of the wolves in the training data were pictured in snowy environments). [80] *This type of tool might likewise serve to highlight whether a system is engaging in "reward hacking" or other problematic behaviours that might otherwise only surface after the system is deployed in the real world. Source: Ribeiro et al. (2016).*

a system is highly unpredictable, no amount of testing will guarantee that it will behave as intended in the field.

A rigorous **verification and validation**[77] process for any autonomous weapon system or complex AI agent will likely depend on gaining insight into why the system operates as it does and what it will do in deployment in order to certify that the system meets the specific requirements and other criteria that were set for it at the beginning of the design process.[78] Equally, this kind of insight will help determine whether the system would meet these requirements in the field to the same degree that it meets them in testing. (Bridging the gap between performance in testing, verification and validation and performance in the field is especially important, and potentially very challenging, when it comes to autonomous systems: for more on this, see page 19).

The **legal review process**, an obligation for all States before fielding new weapons, is a key step that serves to determine whether the intended use of a weapon would, in some or all circumstances, violate the international law obligations applicable to that State. In legal reviews of complex AI systems, this assessment would likely rely on an ability to anticipate a system's performance in the functions and environments for which it is proposed, identify its edge cases and points of failure, model the ways in which it would fail and the effects of those failures, and illuminate any other potential harms that could arise from its use.[79]

The ability to accurately anticipate a system's behaviours could also inform the development of additional doctrine or constraints that would help ensure that any use of the AI system in question complies with international law, in particular IHL, as well as national law and other applicable requirements.[81] For example, if reviewers can determine that a system exhibits a tendency to fail dangerously in response to a particular type of environmental factor, they could require that it shall only be used in tightly constrained environments where that factor is unlikely to be present.

3.2 DURING EMPLOYMENT

Once a system is fielded numerous planning steps, at **various military command levels**, preceding the final decision to employ an AI system[82] serve to determine whether engaging the AI system would facilitate or hinder the execution of the objectives of the operation and could feasibly do so in compliance with IHL and the rules of engagement,[83] anticipate the risks of employment,[84] and develop appropriate parameters and constraints for the system's use to avoid those risks and comply with the relevant rules and laws.[85] These determinations would necessarily be based on some level of understanding of how the system works and, by extension, an assessment of how it is likely to respond to the particular environment at hand.[86] (These factors may be particularly important when it comes to human-out-of-the-loop weapons that lack a communications link with their human operators and cannot be recalled[87]). In short, the more information that decision makers have regarding what the system will do and why it will do it, the better equipped they will be to make the most responsible decision.

While an AI system is actively operating, predictability and understandability are two complementing factors of **robust human-machine interaction**[88] as they enable operators to **precisely calibrate their trust in the system,**[89] thus guiding its appropriate use (or non-use) for every given context.[90] Specifically, to calibrate their trust, operators require some capacity to anticipate or identify system malfunctions; understand how the system is likely to respond to the specific environmental factors at hand; confirm that the system's outputs are not being disproportionately swayed by extraneous factors;[91] and recognize when a system is encountering edge cases or inputs for which it was not designed,[92] has ingested bad data, is being targeted by adversarial attacks[93] or is exhibiting brittleness.[94] In cases where humans are teaming with highly autonomous agents or weapon systems, understandability and predictability may also be key to ensuring that the machine's goals are aligned with those of the human operator.[95]

The need for robust trust calibration applies both to exercising direct control over a system during use (for example, appropriately trusting in the validity of a target that a system is proposing to attack) as well as to employing a fully autonomous weapon system (for example, appropriately trusting whether the employment of the system would be legal or would achieve the desired effects without resulting in failures or undue harm).[96]

Many algorithmic systems present outputs with a confidence score that indicates the likelihood that the output is correct. While it might be tempting to think that this can make systems more understandable and predictable, these scores are usually not sufficient for an operator to precisely calibrate trust, especially if that operator has a minimal technical literacy or if they must calibrate their trust in a short window of time.[97]

3.3 AFTER EMPLOYMENT

The requirements for predictability and understandability do not cease to be relevant after the deployment of the AI weapon system. In **post-use assessment**,[98] organizations seeking to audit or investigate an operation involving AI will desire a level of direct insight into the AI system to aid in determining exactly what happened, and why. This includes determining why the relevant AI agents behaved as they did,[99] and why certain effects resulted from their employment.[100] It also extends to a consideration of whether those responsible for an AI system's use reasonably anticipated the effects of their actions (a legal requirement under IHL), took reasonable steps to avoid undue harms, and followed IHL and rules of engagement in other regards.[101]

Understandability and predictability would also aid in the determination of whether the adverse effects of an AI system's use in a particular instance are likely to emerge again. Such assessments will inform regulatory, doctrinal or technological reforms to prevent future harms.[102]

4. WHAT IS APPROPRIATE PREDICTABILITY AND UNDERSTANDABILITY?

As the previous section illustrates, both predictability and understandability are necessary to some degree at every stage in the development, employment and assessment of military AI systems. But what would constitute an appropriate level and type of predictability and understandability in each of these cases? While this is a complex question that calls for extensive research, it is likely that the exact required level of predictability and understandability for military AI systems will be shaped by the following questions and considerations, among other factors.[103]

- **How critical is the function?** In highly critical roles where the effects of system failure or misuse could be catastrophic, the requirement for understandability and predictability at all stages of development, employment and assessment is likely to be higher.[104] Conversely, for non-critical AI, the requirement for understandability at certain touchpoints of human control, as well as the requirement for technical predictability, may be lower.[105]

- **Are there risks of escalation?** Given that the application of unpredictable black box AI in particularly sensitive contexts (for example, contested border regions) could pose a risk of rapid inadvertent escalation,[106] the requirement for predictability and understandability in such contexts may be higher.

- **What is the form of human control?** Operators can "control" AI in a variety of ways,[107] and requirements for predictability and understandability will vary accordingly. For example, requirements may vary depending on whether the AI is a human-in-, human-on- or human-out-of-the-loop system; whether control is exercised solely through the establishment of parameters before employment or through some form of direct supervision; and whether there is a possibility of recalling or aborting the system.[108]

- **How much operator time and attention is available?**[109] In cases where a human operator only has a brief window to review a system's explanation for a given output, and a high cognitive load[110]—for example, in a case where an AI system identifies a time-sensitive target with only a brief opportunity to strike while a range of other factors are competing for the operator's attention—there may be a different requirement for understandability and predictability than there would be in cases where the operator has the time and mental space to conduct a meticulous review of the system's logic process, past performance and other factors.[111]

- **Can the operator access or understand the input data?**[112] If operators using human-in- or human-on-the-loop AI systems can directly vet the source data of an AI system's input, the requirement for system transparency—or explainability tools —might be lower, assuming that the source data are easy to vet. For example, if an operator can review images of objects that an AI-based automatic targeting system proposes to shoot at, the operator may not require an additional explainability function since they can directly verify the image. However, if the system's proposals derive from, say, a matrix of video, signals intelligence, radar and human intelligence,[113] the operator probably could not reasonably vet all the data, and so system transparency or an explainability function might be necessary to highlight the particularly salient individual inputs.[114]

- **Where is the system being employed?** Given that large, complex adversarial environments are more likely to present systems with inputs for which they have not been trained or tested, a higher level of understandability and predictability may be required in such environments. For example, a fixed aerial defence system operating in a sparse environment where civilian objects and persons are not likely to be present would have different—probably lower—requirements for predictability and understandability than a highly autonomous vehicle operating in a cluttered and adversarial urban environment, especially if civilian objects and persons are likely to be present alongside combatants,[115] or if the status of objects could switch during the battle.[116]

- **Who is the adversary?** Because of the specific challenge posed by adversariality to safe and legal AI system employment,[117] systems that will operate in highly adversarial environments where there is a significant likelihood of subterfuge, data tainting or other AI countermeasures may require higher levels of predictability and understandability so that operators can identify when the system is falling prey to such attacks.[118]

- **(For understandability) Who is the human user or assessor of the system?** Because different individuals have different capacities and motivations for understanding the AI they are interacting with, the **audience** of an AI system is a key determining factor in the requirement for the type and level of understandability.[119] An auditor validating a system before deployment may need an in-depth understanding of the system's technical architecture, while a commander may only want a broad understanding of how the system is likely to respond in a forthcoming operation, and an operator may require a robust mental model to ensure that their trust is always calibrated.[120] (Some groups have begun to define audience categories and their corresponding understandability requirements in the civilian domain.[121])

- **(For understandability) Why is the system being understood or anticipated?** Just as requirements for understandability will vary depending on who is interacting with the system, they will likewise vary according to why those individuals are interacting with the system. For example, appropriate understandability will differ depending on whether a system is being assessed for testing and certification, strategic or operational planning, direct human control or supervision, or after-action assessments and audits.[122]

It should be noted that it is unlikely that all militaries wishing to use AI and autonomous combat technologies will share the same desired thresholds for appropriate understandability or predictability in all contexts. In a high-intensity engagement, for example, operators or other stakeholders may jettison previously rigorous internal requirements for understandability and predictability,[123] especially when operators only have a limited window to understand the AI system and any delays stemming from a requirement for understandability might result in harm—all the more so if stakeholders know that their adversary is using faster, more capable black box systems of their own. In some instances, given the widely held view of the performance-understandability trade-off (see page 10), stakeholders might want to opt for a less inherently transparent system if that system could be more lethal or more precise than a more transparent alternative.[124] Similarly, given that unpredictable activity is a common tactic in warfare to sow uncertainty in the adversary[125] and predictable autonomous systems might be more vulnerable to countermeasures, militaries might seek to employ systems that achieve desired effects through unpredictable actions.[126] These factors may complicate efforts to establish universal baselines of predictability and understandability required for compliance either to existing IHL or to potential new fit-for-purpose rules.

5. MOVING FORWARD

Achieving and assuring these appropriate levels of predictability and understandability in LAWS and other forms of military AI will be a profoundly complex challenge. As discussed earlier, artificially intelligent systems can be inherently unpredictable and unintelligible. These inherent qualities are likely to be compounded by the types of environment that these technologies will be subjected to in conflict. Even the enforcement of blanket bans or restrictions on the use of unpredictable or unintelligible systems would still depend on instruments to measure and test predictability and understandability in a diverse range of complex computer systems—a formidable challenge. A plurality of experts consulted noted that any potential measures to address the black box dilemma will likely implicate new lines of action and inquiry related to **testing**, personnel **training** and **standards**, as well as a closer analysis of the feasibility of **XAI**.

5.1 TESTING, TRAINING AND STANDARDS

5.1.1 Testing

It is impossible to anticipate or understand the behaviours, failures and effects of an AI system operating in conditions for which it has not specifically been tested.[127] As such, one potential option to improve system predictability and understandability is to overhaul testing, evaluation, verification and validation regimes[128] so as to cover the broadest possible range of inputs and environments that the system could foreseeably encounter once fielded. This might include a wide range of adversarial contexts,[129] as well as interactions with other complex systems in the battlespace.[130] For machine learning systems there might additionally be requirements, at the development stage, for the data sets used to train the algorithms to cover the broadest possible range of inputs, with a wide distribution of variables, as well as for "adversarial training" to be conducted, so that the likelihood of systems encountering inputs not covered by their training is reduced.[131] To bolster these safeguards, organizations may additionally need to require that systems only be employed in environments for which they have been specifically and rigorously tested (for example, a computer vision system that has only been tested for clear-weather daytime operations would be prohibited for use in night-time operations or in adverse weather).

That being said, given the inherent operational unpredictability of employing autonomous systems, particularly systems that have a wide decision space or that operate in complex and dynamic environments (see page 7), it will be challenging to certify that a system will respond safely or appropriately to every possible input and condition it might encounter in deployment.[132] Certain edge cases may only arise from a very specific and unforeseeable set of interacting circumstances. Testing a system against all such potential combinations of circumstances that might give rise to all of its possible failures could be extremely challenging.[133] Furthermore, the possible outcomes of a failure, or even of the interactions between the AI system and other complex systems, are potentially far more diverse and difficult to model[134] than the outcomes of a malfunction in, say, a missile on a ballistic trajectory—again, especially if the AI system is highly autonomous and the proposed deployment environment is complex.[135]

It has therefore yet to be established whether traditional methods for assessing complex weapons would be sufficient to provide a clear view on the extent to which a highly autonomous system is likely to behave as intended or fail as anticipated.[136] Even newer testing techniques developed specifically for machine learning systems would likely struggle to cover all the potential edge cases that may arise in a very complex environment such as a battlefield.[137] This could be particularly challenging when it comes to learning AI systems that continuously tweak their parameters to improve their performance. While such systems could offer significant gains by progressively "fitting" their statistical models to the deployed environment, thus potentially reducing brittleness, such systems could also potentially acquire behaviours that have not been tested and certified.[138]

To address these concerns, some observers propose a recursive testing, evaluation, verification and validation process, where systems (both active and non-active learning) are continually tested and certified to ensure that they continue to meet their safety and legal requirements. Where necessary, these systems can be updated in response to feedback from the field, assuming that it can be certified that these updates would not, in turn, generate new failures.[139] Given that both understandability and predictability centre on the human subject understanding or anticipating a system, it is likely that testing and verification of these attributes in AI systems would need to include rigorous human evaluations[140] to certify that the model in question exhibits the desired characteristics when operating with or under the specific type of user for whom it is designed.[141] To be fully accurate, these evaluations may need to involve the actual types of individual who will be anticipating or interpreting the behaviours of the system in real life.[142]

More broadly, it has not yet been determined **how predictability and understandability can be consistently and reliably measured, since both are influenced by complex variables (like environmental factors) and fuzzy characteristics such as "human capacity for understanding"**. In advance of the establishment of evaluation programmes, it is therefore widely agreed that extensive further research will likely be necessary to establish viable measurable criteria for grading understandability and predictability.[143]

5.1.1 Training

An oft-cited option to enhance individuals' capacity for understanding AI systems is rigorous technical and operational training.[144]

Given the complex computer science underlying all forms of AI, technical literacy training could potentially enable individuals to better grasp how systems actually function. The resulting technical understanding may be more effective than understanding based on explainability tools that only provide an approximate abstraction of the system's logic (see page 22) or the advice of AI interpreters who "translate" concepts for lay users.[145]

Operational training could potentially strengthen human subjects' mental models of the systems they interact with and ensure that they do not encounter situations for which they do not have any understanding of the systems' likely responses.

Such technical and operational training regimes may have to be applied for every human subject that will be interacting with AI systems[146] at every stage of development, employment and assessment—including senior commanders determining whether to use such systems, legal counsellors assessing how the systems could be used within the limits of the law, and auditors investigating incidents after the fact. In all of these cases, such training would have to be expansive, detailed and tailored to match the complexity of the AI systems in question and the varying required levels of understanding dictated by the role.[147]

5.1.2 Standards

Implementing and assuring appropriate predictability and understandability would likely hinge on extensive new technical research as well as a significant overhaul of current doctrine.[148] One potential measure that could provide guidance and consistency in these efforts is the creation of standards. A number of efforts are currently ongoing to establish standards and guidance for enshrining, measuring and certifying predictability and understandability in civilian AI systems,[149] and it has been suggested that these kinds of standards may provide a robust template for standards for military AI.[150]

However, the implementation of such standards rests on several unanswered questions. For one, it is possible that certain AI technologies are "pre-standard". This is to say, because these technologies are still immature they are likely to continue evolving in ways that could render any overly premature standards obsolete.[151] Additionally, measures to apply civilian standards to military AI may have to account for the inherent differences between the two domains, namely in terms of the adversariality of conflict as well as the requirement of secrecy.[152] More research is therefore needed on the feasibility of standards specifically for LAWS and other forms of military AI.

5.2 A TECHNICAL APPROACH: EXPLAINABLE AI

5.2.1 What is Explainable AI?

As concerns about understandability have come to the fore, there has been growing interest and investment in efforts to address the challenge by technical means. This field of research is known as Explainable AI.

Within the field of XAI, there are two essential (and very different) approaches to engendering understandability. One approach seeks to build, from the ground up, high-performance AI systems that are intrinsically interpretable; such systems are sometimes referred to as "transparent models" or "self-explainable" models.[153] The second approach seeks to build add-on tools that can "explain" or "translate" unintelligible AI systems. Such explanations can take a variety of forms, including:[154]

- **Verbal descriptions** that explain how a particular output was achieved. For example, a tool that informs the user that "the system identified the object as an aircraft because of its colour, its radar cross section, and its altitude and speed".
- **Visualizations** that highlight features of the input data that were particularly definitive for the resulting output (for example, a technique known as "saliency mapping").
- **Counterfactual or contrastive explanations** that illustrate why the system generated the given output and not something else.
- **Approximate models** and other abstractions, which essentially replicate the black box system's process with a simplified interpretable model.

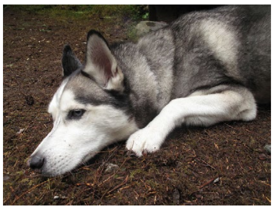
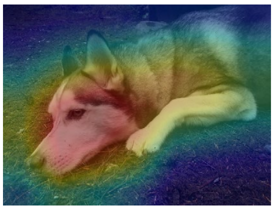
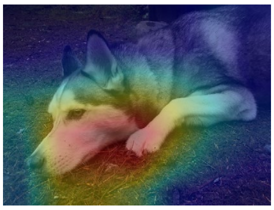
> XAI draws on knowledge from a range of non-technical fields like psychology, human factors and other social sciences. For example, to build a strong explanation tool, XAI researchers may need to first answer the very philosophical question, *What even is an explanation?*[155]

5.2.2 Challenges of Explainable AI

To be truly useful, an explanation must be accurate, clear and meaningful.[156] While XAI is often described as a present-day solution to the black box dilemma, the challenges of building XAI that meet all these and other criteria are manifold.

For one, given that requirements for understandability will vary widely according to a broad range of factors (see page 17) and XAI tools vary in terms of the type of understanding they engender, it is difficult for researchers to determine what explanatory information is necessarily relevant or irrelevant in any single given application.[157] In testing, operators might require detailed technical XAI-generated information; in live operations, where the operator is likely to contend with many intensive demands on their attention, a detailed explanation will not be of any value[158] or might even do more harm than good.[159] Therefore, each type of user may require different types of explanations.

FIGURE 3

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Sometimes, explainability tools can be wrong. In this example, a "saliency map" explainability tool that shows users which parts of an image were most influential in the AI system's analysis of that image provides a nearly identical explanation for both a correct identification (centre) and an incorrect one (right). Source: Chaofan Chen, and Rudin (2019).

Add-on explainability tools may not be a viable proposition for autonomous weapons that do not maintain a regular communications link with the operators (for example, an autonomous drone). Additionally, if such systems are non-retrievable—for example, an autonomous missile system—their explanations would likewise be unavailable[171] for after-action assessments, audits and investigations.

Compounding this issue is the difficulty of testing the practical value of XAI approaches through human evaluations.[160] Those XAI studies that do include human evaluations[161] have not generally tested subjects in scenarios that would be approximate to warfighting environments, where the stakes of decisions are very high, the margin for error is low and the cognitive burden on operators is crushing.[162]

Furthermore, explainability tools can only ever provide an approximation of the actual opaque AI system.[163] As such, the explanations are not always sufficiently detailed to provide insight into the correctness of the AI system's output[164] or, worse, explanations can be incorrect.[165] And when an explanation is found to be incorrect, the user's trust in the system is likely to drop considerably—undermining effective trust calibration.[166]

A converse challenge is the tendency of explanation tools to engender over-trust in AI systems.[167] Systems that provide an eminently understandable explanation of a highly complex system may in fact offer very little insight into whether the AI system is right or wrong (see Figure 4), especially if these explanations match the user's expectations or if the user is biased in favour of the XAI tool.[168] For these reasons, some have proposed that inherently understandable AI architectures are preferable to opaque systems with explainability add-ons, especially when both options yield similar performance.[169]

For these reasons, XAI remains a bleeding-edge research problem. Even much of the most advanced technical work in this realm is still only foundational.[170] **It would therefore be potentially problematic to build policy or norms on the assumption that reliable, replicable understandability for complex AI in critical roles can be achieved by technical means alone in either the short or medium term.** It may be safer to assume that a lack of understandability could continue to be an inherent aspect of complex AI in all the roles for which it is being considered and that technical explainability measures will at best serve to complement non-technical approaches.

CONCLUSION: FIVE AVENUES FOR ACTION

This study has sought to establish a baseline of common knowledge related to understandability and predictability in military applications of AI. It points to a variety of avenues for action by the policy community, the technical community and military organizations:

1. **Adopt a common taxonomy and framing of predictability and understandability**. There are various different senses of both predictability and understandability, and the many related— sometimes overlapping—terms and ideas related to this issue remain poorly defined. This hinders precise, effective dialogue.

2. **Explore non-military initiatives on AI understandability and predictability**. Issues related to understandability and predictability have been explored extensively in the civilian sector. A number of organizations have even begun to take substantive steps to address these challenges. While the civilian domain and the battlefield differ in many crucial respects, some of the approaches currently under consideration in critical domains such as transportation, medicine, finance and security could potentially serve as object lessons for the ongoing debate on LAWS and other forms of military AI.

3. **Study the factors that determine appropriate levels of understandability and predictability.** As described in Section 4, a wide range of factors determine the appropriate level of understandability and predictability in any given instance of development, employment or after-action assessment. A common formulation of these factors could serve as a foundation for debate on the adoption and implementation of potential frameworks, norms or standards for human-machine interaction and could highlight areas of shared understanding or divergence on such instruments and measures.

4. **Develop standardized metrics to grade predictability and understandability.** While it is broadly agreed that understandability and predictability are fundamental to the proper employment of AI in combat operations, more work is needed on how to reliably measure these characteristics in military systems.

5. **Assess the viability of training and testing regimes that can engender robust AI understanding and account for AI unpredictability.** Given the centrality of testing and training for assuring understandability and predictability in AI, and given the challenges of testing and training described in Section 5, publicly shared assessments of the viability of such training and testing techniques could prove useful for guiding the policy debate.

All the lines of action listed here will require close collaboration. Policy must be grounded in technical expertise, technical solutions must align with legal requirements, and national doctrine must match shared norms and principles. While this is true of the ongoing debate over LAWS and military AI in general, it may be particularly true of understandability and predictability. The black box dilemma stands squarely at the intersection of technical, normative and doctrinal considerations. This report therefore illustrates the urgent need for a diverse, rigorously interdisciplinary, cross-cutting dialogue between all stakeholders.

ENDNOTES

1. According to a study in *Nature Machine Intelligence*, 73 of 84 international artificial intelligence ethics documents
vaunt the principle of "transparency", which includes predictability and understandability: Jobin et al. (2019, 391).
2. For example, GGE on LAWS (2019, 3).
3. The terms "AI system" and "autonomous system" are not interchangeable, but they do overlap. For the purposes of this report, autonomous systems usually achieve autonomy through an AI function, but not all systems that include an AI function can necessarily be described as autonomous.
4. Interviews with Raja Chatila, 26–27 May 2020; interview with Pascale Fung, 3 June 2020.
5. Interviews with Raja Chatila, 26–27 May 2020; for further discussion of these terms in relation to weapon systems, see UNIDIR (2017, 13–14).
6. Boulanin (2019, 20, 133); Defense Innovation Board (2019, 16). For a detailed discussion of autonomous system failures, see UNIDIR (2016).
7. Defense Innovation Board (2019, 16); Heaven (2019); IEEE (2017b, 128).
8. After all, the key proposed benefit of autonomous systems is that they will be capable of achieving goals that humans cannot achieve on their own and of navigating situations of which the operators do not have specific prior knowledge. Interview with Liran Antebi, 8 June 2020; interview with Pascale Fung, 3 June 2020; Boulanin & Verbruggen (2017, 6); Williams (2015, 33, 57).
9. Interview with Geoffrey M. Schaefer, 4 June 2020; interview with Lindsey Sheppard, 2 June 2020; UNIDIR (2017, 12–13); Yampolskiy (2019).
10. IEEE (2017b, 128).
11. For example, in its 2018 position paper to the Group of Government Experts on LAWS, the Chinese delegation listed the ability to "expand its functions and capabilities in a way exceeding human expectations" as one of the five defining characteristics of LAWS: China (2018, 1).
12. For a similar illustration, see ICRC (2019b, 11).
13. Docherty (2020); ICRC (2019b, 10–11); IEEE (2017b).
14. Chan (2017).
15. Yampolskiy (2019, 2).
16. Interviews with Raja Chatila, 26–27 May 2020; interview with Pascale Fung, 3 June 2020; Ribeiro et al. (2016, 100).
17. Amodei et al. (2016, 8). For a brief overview of AI approaches, see Fumo (2017).
18. Interview with anonymous expert, 2 June 2020.
19. Boulanin & Verbruggen (2017, 70).
20. Amodei et al. (2016, 16); Pontin (2018).
21. Interview with anonymous expert, 28 May 2020; Wu et al. (2019).
22. For example, if a computer vision tool that was trained on a data set of images of dogs were given an image of a helicopter.
23. For example, if a facial recognition system trained on a data set of predominantly young faces were deployed on a real population of people with a broad range of ages. Quiñonero-Candela et al. (2009); Stewart (2019).
24. Interview with Pascale Fung, 3 June 2020; Goussac (2019).
25. Crootof (2016, 1350); CRS (2019, 17); ICRC (2019b, 2); Scharre (2020).
26. US Department of Defense (2017, 6). For a broader discussion of emergent behaviours in complex systems, particularly multi-agent systems like autonomous swarms, see Ilachinski (2017).
27. Interview with Patrick Bezombes, 26 May 2020; Boulanin & Verbruggen (2017, 17); Collopy et al. (2020, 48); ICRC (2019b, 16); UNIDIR (2017, 10).
28. Interview with anonymous expert, 2 June 2020; Boulanin et al. (2020, 12).
29. Interview with Henrik Røboe Dam, 27 May 2020; interview with Liran Antebi, 8 June 2020.
30. ICRC (2019b, 12).
31. Interview with Cynthia Rudin, 3 June 2020; Sessions & Valtorta (2006, 485).
32. Comiter (2019); Feldman et al. (2019, 4); O'Sullivan (2019). Recent research has demonstrated that certain machine learning systems are capable of deception, which could also make such systems far more unpredictable: Roff (2020).
33. Interviews with Raja Chatila, 26–27 May 2020; interview with Kerstin Vignard, 9 June 2020; IEEE (2017b, 128); Ilachinski (2017). For more on swarm control, see Ekelhof & Persi Paoli (2020a).
34. Barredo Arrieta et al. (2019, 5); Bhatt et al. (2020, 648); Carvalho et al. (2019, 3–7). For a beginner's overview on the topic, see Schmelzer (2020).
35. Kuang (2017); UNIDIR (2017, 7).
36. No single universal taxonomy for types of understandability has emerged. See for example, Arya et al. (2019); Barredo Arrieta et al. (2019, 10–11).

37. Given that the core concepts and terms in this area can be fuzzy and poorly defined across the literature, it is important to be aware of these different usages when navigating the discourse: Lipton (2016, 1). For discussions of these varying definitions, see Barredo Arrieta et al. (2019, 5); Bhatt et al. (2020, 649); Castelluccia & Le Métayer (2019, 27–29); Clinciu & Hastie (2019).
38. Bornstein (2016); Leslie (2019, 43); UNIDIR (2017, 7, 11–12).
39. Dickson (2020).
40. Interviews with Raja Chatila, 26–27 May 2020; interview with Pascale Fung, 3 June 2020.
41. This example is drawn from an experiment by Ribeiro et al. (2016). For another example of this phenomenon and why it can be problematic in critical applications of AI, see Caruana et al. (2015, 1721–22); Kuang (2017). For another, more recent example, see Brendel (2019).
42. Amodei et al. (2016); Krakovna et al. (2020).
43. This view was supported almost unanimously by the subject matter experts interviewed for this study.
44. Castelvecchi (2016); Lipton (2016, 2).
45. Castelluccia & Le Métayer (2019, IV); Mittelstadt et al. (2019, 280).
46. To use the example of the husky/wolf identification system, a local explanation would explain why the system identified one particular dog as a wolf, whereas the global explanation would explain how the system identifies wolves in general. Interview with anonymous expert, 2 June 2020; Diop (2018).
47. Interview with Cynthia Rudin, 3 June 2020; Schmelzer (2020).
48. Boulanin et al. (2020, 20).
49. Interview with anonymous expert, 28 May 2020; LeCun (2020).
50. Stubbs et al. (2007, 47).
51. See, for example, Barredo Arrieta et al. (2019, 30); Caruana et al. (2015, 1721); DARPA (2016, 7). Interview with Geoffrey M. Schaefer, 4 June 2020.
52. DARPA (2016, 5).
53. Ilachinski (2017, 182).
54. Interview with anonymous expert, 29 May 2020.
55. Barredo Arrieta et al. (2019, 44).
56. Roff & Danks (2018, 3). While this is a widely held belief in the field of computer science, some have challenged this paradox, pointing out that the trade-off between understandability and performance does not always, or even regularly, hold true. See Leslie (2019, 44); Rudin (2019).
57. Slayton (2020).
58. Interview with anonymous expert, 9 June 2020; Barredo Arrieta et al. (2019, 7).
59. Interview with Kerstin Vignard, 9 June 2020. For more on AI and adversariality, see Comiter (2019); O'Sullivan (2019); Van den Bosch & Bronkhorst (2018).
60. In particular, Principles b through e: GGE on LAWS (2019, annex IV).
61. As yet, there is no common understanding of the exact meaning of each of these terms. For a discussion of predictability and understandability in teaming operations, see Joe et al. (2014).
62. CCW (2019, 10).
63. Interview with Patrick Bezombes, 26 May 2020; interview with Henrik Røboe Dam, 27 May 2020.
64. Interview with Geoffrey M. Schaefer, 4 June 2020.
65. As several experts pointed out, no stakeholder who is ultimately responsible for a system's effects would want to deploy a system that is highly unpredictable or uninterpretable. Interview with Patrick Bezombes, 26 May 2020; interview with Henrik Røboe Dam, 27 May 2020; interview with anonymous expert, 28 May 2020; interview with anonymous expert, 29 May 2020; interview with anonymous expert, 2 June 2020.
66. The international community has yet to reach a common understanding of "meaningful" in this context.
67. For discussions of predictability and understandability as key elements of meaningful human control, see Chengeta (2017); Docherty (2020); ICRC (2018, 2); Roff & Moyes (2016); UNIDIR (2014, 5).
68. Interview with anonymous expert, 9 June 2020; Article 36 (2016); Boulanin et al. (2020, 12).
69. GGE on LAWS (2018) defines a framework of five "touch points in the human-machine interface", which span from the testing and evaluation stage through to the employment and review stage. The notion that predictability and understandability are required throughout the development and employment of AI also has currency in the civilian realm: see Doshi-Velez & Kim (2017, 2–4); Leslie (2019, 37).
70. For example, the role of understandability could be particularly weighty in lethality-enabling autonomous weapons systems—systems that are not in themselves "on the trigger" but that directly support the employment of force: Holland Michel (2020).
71. Bhatt et al. (2020, 649).
72. The issue of algorithmic bias has been raised as a significant concern with respect to the application of AI in critical roles. For more on types of bias and their implications in autonomous systems, see Danks & London (2017).
73. Caruana et al. (2015).
74. For a discussion of some of the most common data issues as they relate to military applications of AI, see

Kostopoulos (2018).
75. Krakovna et al. (2020).
76. Brundage et al. (2020, 26).
77. For an overview of the verification and validation process, see IEEE (2017a).
78. Boulanin & Verbruggen (2017, 70); Hagström (2019, 37); Ilachinski (2017, 199–209). For a technical survey of approaches to verification and validation for autonomous systems, see Ingrand (2019). To embed understandability at the earliest stage of AI development, some groups, particularly in the civilian domain, have suggested defining explicit explainability and predictability criteria in AI procurement requirements. For example, Ethical AI Institute (2020).
79. For a detailed primer on the kinds of negative side effects that can arise from machine learning systems in particular, see Amodei et al. (2016).
80. Ribeiro et al. (2016, 100). See also UNIDIR (2017, 7–8).
81. Davison (2017, 9–10).
82. Military decisions to use force run through a complex multilayered process, taking into account the input of multiple agents within the chain of command. For a full description of this process, see Ekelhof & Persi Paoli (2020b).
83. Interview with David Barnes, 4 June 2020; interview with Thompson Chengeta, 26 May 2020; interview with Henrik Røboe Dam, 27 May 2020; Lawand (2020).
84. Interview with David Barnes, 4 June 2020.
85. A process sometimes known as “bounding”: Haugh et al. (2018); see also ICRC (2019b, 12).
86. Interview with Henrik Røboe Dam, 27 May 2020.
87. Interview with Cynthia Rudin, 3 June 2020; interview with Kerstin Vignard, 9 June 2020.
88. GGE on LAWS (2019, 3–4).
89. Trust calibration refers to the ability to moderate one’s trust in an AI agent in response to contextual factors—for example, the input data quality—that may affect that system’s ability to perform as intended. There is extensive literature on the relationship between understandability and/or predictability and trust. See CRS (2019, 31); Devitt (2018); Dietvorst et al. (2016); Lewis et al. (2018); Morgan et al. (2020, 36); Wang et al. (2016). For a discussion on some of the fundamental problematics of trust, see Hoffman (2017).
90. Boulanin et al. (2020, 20); Hawley (2017, 12); Kiernan (2015, 4); Lewis et al. (2018); Roff & Danks (2018); Van den Bosch & Bronkhorst (2018, 9).
91. Doshi-Velez & Kim (2017, 2); Ribeiro et al. (2016, 98).
92. Phillips et al. (2020, 4) describe this concept of "knowledge limits" as a fundamental principle of explainability.
93. Baksh (2020).
94. Dix et al. (2003, 163); Leslie (2019, 32).
95. CRS (2019, 32); Ilachinski (2017, 186–87).
96. One anonymous expert (interviewed 29 May 2020) explained that if an AI system produces an erroneous output but provides a reasonable explanation for that output, the operator’s trust in that system will not be undermined; in other words, the operator will ignore the specific output but will continue to trust the system overall.
97. See for example, Galyardt (2018); Miller (2019, 6); Snoek & Nado (2020).
98. For more on this touchpoint of human control, see GGE on LAWS (2018, 15).
99. Interview with Thompson Chengeta, 26 May 2020. In other forums, stakeholders and experts have argued that unpredictability creates an "accountability gap" or "accountability confusion" because it is hard to hold an actor liable if one cannot prove knowledge or intent. See Chengeta (2015); HRW & IHRC (2015); ICRC (2014, 8).
100. One anonymous expert (interviewed 9 June 2020) cautioned that an AI system’s explanation alone should not be used to determine whether a particular action involving that system was legal, since it is widely agreed that machines cannot be held accountable for actions in warfare: ultimate responsibility rests with the human operator.
101. See GGE on LAWS Guiding Principle c in GGE on LAWS (2019, 13).
102. For a civilian sector parallel, see the concepts of “answerability” and “auditability” in Leslie (2019, 24).
103. Interview with anonymous expert, 29 May 2020; Brundage et al. (2020, 26).
104. Interview with Liran Antebi, 8 June 2020; interviews with Raja Chatila, 26–27 May 2020; interview with Henrik Røboe Dam, 27 May 2020; interview with Cynthia Rudin, 3 June 2020; interview with Lindsey Sheppard, 2 June 2020; Brundage et al. (2020, 26); Defense Innovation Board (2019, 38).
105. Interview with Cynthia Rudin, 3 June 2020; interview with anonymous expert, 2 June 2020; LeCun (2020); Simonite (2018).
106. Interview with Liran Antebi, 8 June 2020; Garcia (2019); Huh Wong et al. (2020, 66); Rickli (2019, 93).
107. For more on the various types of human control of autonomous systems, see Musco Eklund (2020, 13–27).
108. Interview with Cynthia Rudin, 3 June 2020; interview with Geoffrey M. Schaefer, 4 June 2020; interview with Kerstin Vignard, 9 June 2020.
109. Interview with Henrik Røboe Dam, 27 May 2020; interview with Geoffrey M. Schaefer, 4 June 2020; interview

with Lindsey Sheppard, 2 June 2020; interview with anonymous expert, 29 May 2020.

110. Boulanin et al. (2020, 20).

111. Doshi-Velez & Kim (2017, 8); Phillips et al. (2020, 5–6).

112. Interview with anonymous expert, 29 May 2020.

113. Interview with Cynthia Rudin, 3 June 2020; interview with anonymous expert, 29 May 2020.

114. Saliency, or relevance, mapping is a favoured explainability technique for precisely this reason. See for example Petsiuk et al. (2018).

115. Interview with Liran Antebi, 8 June 2020; interview with Henrik Røboe Dam, 27 May 2020.

116. Interview with Thompson Chengeta, 26 May 2020. For instance, when a combatant signals a surrender and thus takes on protected status. For a specific discussion of surrender recognition as it relates to LAWS, see Sparrow (2015); for a general discussion of surrender, see Buchan (2018, 11).

117. Danks (2020); Feldman et al. (2019, 4); O'Sullivan (2019). For a broad overview of AI vulnerabilities, see Comiter (2019).

118. Interview with Liran Antebi, 8 June 2020. According to a senior official at the US Cybersecurity and Infrastructure Security Agency, the threat of adversarial attacks is a primary driver of understandability requirements for the agency's AI systems. See Baksh (2020).

119. This was a view expressed by all experts interviewed for this study. See also Bhatt et al. (2020, 650); Ribera & Lapedriza (2019, 4); The Royal Society (2019,19).

120. Interview with Pascale Fung, 3 June 2020; interview with anonymous experts, 28–29 May 2020.

121. Bhatt et al. (2020, 650) propose four audience categories for the civilian AI realm: executives, engineers, end users, and other stakeholders. Ribera & Lapedriza (2019, 4) propose three tiers: developers and AI researchers, domain experts, and lay users.

122. Interview with Raja Chatila, 26–27 May 2020; interview with Pascale Fung, 3 June 2020. Phillips et al. (2020, 4) define five purposes of AI explanations—including to directly benefit the operator, to support audits and compliance, and to guide further system development—and note the different types of understandability appropriate for each. For further detail, see Bhatt et al. (2020).

123. Interview with Kerstin Vignard, 9 June 2020.

124. Interview with Liran Antebi, 8 June 2020.

125. Greene (2006, 440–41).

126. Hagström (2019, 35).

127. Goussac (2019).

128. For a broad overview of approaches to tailor such regimes to autonomous systems, see Haugh et al. (2018). For a discussion of the challenges posed by AI systems to existing regimes, see Boulanin & Verbruggen (2017, 70); Defense Innovation Board (2019, 16, 66); Hagström (2019); Hall (2017); Hawley (2017, 9); ICRC (2019b, 18–19); Trumbull (2020, 29–30).

129. Interview with Pascale Fung, 3 June 2020; Davison (2017).

130. Interview with David Barnes, 4 June 2020.

131. Interview with anonymous expert, 2 June 2020; Bhatt et al. (2020, 651).

132. Hall (2017); Hawley (2017, 9); ICRC (2019b, 18–19); Trumbull (2020, 29–30).

133. Koopman & Wagner (2016).

134. Boulanin & Verbruggen (2017, 70); Defense Innovation Board (2019, 16, 66).

135. Interview with anonymous expert, 28 May 2020; Goussac (2019).

136. Interview with Lindsey Sheppard, 2 June 2020; UNIDIR (2016, 8). For a discussion of this issue as it relates to validation and verification, see Bagchi et al. (2020, 5); Ilachinski (2017, 199–209); Martin et al. (2019, 61).

137. Bagchi et al. (2020, 5–6).

138. Collopy et al. (2020, 48); Lewis (2018). One expert, Patrick Bezombes (interviewed 26 May 2020), suggested that the use of active learning systems in certain critical applications should potentially be banned for these reasons.

139. Interview with David Barnes, 4 June 2020; interview with Lindsey Sheppard, 2 June 2020; interview with anonymous expert, 2 June 2020; Morgan et al. (2020, 45); Schmelzer (2020).

140. Interview with anonymous expert, 29 May 2020; Lage et al. (2019, 1–2).

141. For example, Herman (2017) has argued that poorly designed evaluations run the risk of models being certified on the basis of persuasiveness rather than descriptiveness.

142. Interview with Pascale Fung, 3 June 2020; interview with Cynthia Rudin, 3 June 2020.

143. Barredo Arrieta et al. (2019, 30); Brundage et al. (2020, 26–27); Lage et al. (2019, 4–7).

144. Interview with Patrick Bezombes, 26 May 2020; interview with Henrik Røboe Dam, 27 May 2020; interview with Lindsey Sheppard, 2 June 2020; interview with anonymous expert, 29 May 2020; Boulanin et al. (2020, 30); Defense Innovation Board (2019, 28, 33, 42); NSCAI (2020, 30).

145. Some observers have proposed AI "liaisons" or management teams that could operate in this capacity in lieu of explainability tools—see Margulies (2016); Roff & Danks (2018, 13)—though it is acknowledged that such an arrangement could introduce a new point of failure into what is already a complex arrangement.

146. Fountaine et al. (2019).

147. Hawley (2017, 9–10) includes specific recommendations for training criteria to build robust mental modelling, including understanding of the system and understanding of edge cases.
148. Interview with anonymous expert, 28 May 2020.
149. Standardization initiatives have included projects at the US National Institute of Standards and Technology and the International Organization for Standardization: NIST (2019); ISO (2017). Leslie (2019, 44) includes a guide for designing and implementing understandable AI in the civilian realm. See also Brundage et al. (2020).
150. Interview with Patrick Bezombes, 26 May 2020; interviews with Raja Chatila, 26–27 May 2020; Daiki (2020).
151. Interview with anonymous expert, 2 June 2020.
152. Interview with anonymous expert, 9 June 2020.
153. Phillips et al. (2020, 9).
154. For broad overviews of these explanation methods, see Biran & Cotton (2017); Mittelstadt et al. (2019, 280); Rai (2020). For a lengthier, more detailed rundown of approaches, see Barredo Arrieta et al. (2019, 10); Bhatt et al. (2020, 651-55); Mueller et al. (2019); Sokol & Flach (2020); Tjoa & Guan (2020, 2–12); Zhang & Zhu (2018).
155. Interview with anonymous expert, 29 May 2020; Miller (2019, 5).
156. For a full discussion of such proposed characteristics, see for example Barredo Arrieta et al. (2019, 6); Brundage et al. (2020, 26–27); Carvalho et al. (2019, 23); Hoffman et al. (2019); Phillips et al. (2020). For a discussion of how to test these characteristics, see Kaur et al. (2020, 2); Ribera & Lapedriza (2019, 4).
157. Interview with anonymous expert, 29 May 2020.
158. Interview with Liran Antebi, 8 June 2020.
159. Interview with anonymous expert, 29 May 2020.
160. Kaur et al. (2020, 1–2); Lage et al. (2019, 1–2).
161. A list of studies that have included human evaluations is available in Mueller et al. (2019, 170). Doshi-Velez & Kim (2017) provide a framework for different types of human evaluation, albeit for civilian, AI applications.
162. Poursabzi-Sangdeh et al. (2018) found that in complex operations some explainability tools actually hampered operators' ability to detect when an autonomous system made an error, since the tool increased their cognitive load.
163. Mittelstadt et al. (2019, 286); Rudin (2019, 207).
164. For an explanation to be a perfect representation of the system it is explaining, it must necessarily be as complex as that system, and therefore it would not be understandable: Yampolskiy (2020).
165. Not only can explanations be incorrect owing to internal failures or faults, they may also be susceptible to adversarial attack. See for instance Lakkaraju & Bastani (2019), as discussed in Phillips et al. (2020, 12).
166. Okamura & Yamada (2020). The phenomenon of under-trust is sometimes referred to as "automation distrust bias": Leslie (2019, 21).
167. Heaven (2020); Kaur et al. (2020, 2).
168. Interviews with Raja Chatila, 26–27 May 2020; Kaur et al. (2020, 5–6).
169. Interview with Cynthia Rudin, 3 June 2020; Ferrini (2019); Heaven (2020); Rudin (2019).
170. Interview with anonymous expert, 29 May 2020; Schmelzer (2019).
171. Interview with Cynthia Rudin, 3 June 2020.

BIBLIOGRAPHY

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané. 2016. 'Concrete Problems in AI Safety.' arXiv:1606.06565v2, 25 July. As of 8 September 2020: https://arxiv.org/pdf/1606.06565v2.pdf

Antebi, Liran, & Gil Baram. 2020. 'Cyber and Artificial Intelligence – Technological Trends and National Challenges.' *Cyber, Intelligence, and Security* 4 (1), March 2020: 131–47.

Article 36. 2016. *Key Elements of Meaningful Human Control*. London: Article 36. As of 18 June 2020: www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf

Arya, Vijay, Rachel K.E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei & Yunfeng Zhang. 2019. 'One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.' arXiv:1909.03012v2, 14 September. As of 23 June 2020: https://arxiv.org/pdf/1909.03012.pdf

Ashmore, Rob, Radu Calinescu & Colin Paterson. 2019. 'Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenge.' arXiv:1905.04223, 10 May. As of 27 July 2020: https://arxiv.org/pdf/1905.04223.pdf

Bagchi, Saurabh, Vaneet Aggarwal, Somali Chaterji, Fred Douglis, Aly El Gamal, Jiawei Han, Brian J. Henz, Hank Hoffmann, Suman Jana, Milind Kulkarni, Felix Xiaozhu Lin, Karen Marais, Prateek Mittal, Shaoshuai Mou, Xiaokang Qiu & Gesualdo Scutari. 2020. 'Grand Challenges in Resilience: Autonomous System Resilience through Design and Runtime Measures.' *IEEE Open Journal of the Computer Society*. doi: 10.1109/OJCS.2020.3006807

Baksh, Mariam. 2020. 'Artificial Intelligence Systems Will Need to Have Certification, CISA Official Says.' NextGov, 1 July. As of 2 July 2020: https://www.nextgov.com/cybersecurity/2020/07/artificial-intelligence-systems-will-need-have-certification-cisa-official-says/166600

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Sera, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopeza, Daniel Molinag, Richard Benjaminsh, Raja Chatila & Francisco Herrera. 2019. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.' Preprint submitted to *Information Fusion*, 58: 30 December.

Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M.F. Moura & Peter Eckersley. 2020. 'Explainable Machine Learning in Deployment.' In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*: 648–57. New York: Association for Computing Machinery.

Biran, Or, & Courtenay Cotton. 2017. 'Explanation and Justification in Machine Learning: A Survey.' *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*. Melbourne, 20 August 2017.

Bornstein, Aaron M. 2016. 'Is Artificial Intelligence Permanently Inscrutable?' Nautilus, 1 September. As of 2 June 2020: nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable

Boulanin, Vincent. 2019. 'Artificial Intelligence: A Primer.' In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. 1, *Euro-Atlantic Perspectives*, edited by Vincent Boulanin, 13–25. Stockholm: Stockholm International Peace Research Institute.

Boulanin, Vincent, Neil Davison, Netta Goussac & Moa Peldán Carlsson. 2020. *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*. Stockholm: Stockholm International Peace Research Institute.

Boulanin, Vincent, & Maaike Verbruggen. 2017. *Mapping the Development of Autonomy in Weapon Systems*. Stockholm: Stockholm International Peace Research Institute.

Brendel, Wieland. 2019. 'Neural Networks Seem to Follow a Puzzlingly Simple Strategy to Classify Images.' Bethgelab, 6 February. As of 15 July 2020: https://medium.com/bethgelab/neural-networks-seem-to-follow-a-puzzlingly-simple-strategy-to-classify-images-f4229317261f

Brundage, Miles, et al. 2020. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.' arXiv:2004.07213, 15 April. As of 8 September 2020: https://arxiv.org/pdf/2004.07213.pdf

Buchan, Russel. 2018. 'The Rule of Surrender in International Humanitarian Law.' *Israel Law Review* 51 (1): 3–27. doi:10.1017/S0021223717000279

Caruana, Rich, Yin Lou, Johannes Kehrke, Paul Koch, Marc Sturm & Noémie Elhadad. 2015. ‘Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission.’ In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 15)*, 1721–30. New York: Association for Computing Machinery.

Carvalho, Diogo V., Eduardo M. Pereira & Jaime S. Cardoso. 2019. ‘Machine Learning Interpretability: A Survey on Methods and Metrics.’ *Electronics* 8 (8): 832. doi:10.3390/electronics8080832

Castelluccia, Claude, & Daniel Le Métayer. 2019. *Understanding Algorithmic Decision-Making: Opportunities and Challenges*. Brussels: European Parliamentary Research Service.

Castelvecchi, David. 2016. ‘Can We Open the Black Box of AI?’ *Nature*, 5 October. As of 23 May 2020: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

Chan, Dawn. 2017. ‘The AI That Has Nothing to Learn from Humans.’ *The Atlantic*, 20 October. As of 8 July 2020: https://www.theatlantic.com/technology/archive/2017/10/alphago-zero-the-ai-that-taught-itself-go/543450

Chengeta, Thompson. 2015. ‘Accountability Gap, Autonomous Weapon Systems and Modes of Responsibility in International Law.’ SSRN, 31 March 2016. doi:10.2139/ssrn.2755211

———. 2017. ‘Defining the Emerging Notion of “Meaningful Human Control” in Weapon Systems.’ *International Law and Politics* (49): 833–90. As of 8 June 2020: https://nyujilp.org/wp-content/uploads/2010/06/NYI303.pdf

China. 2018. *Position Paper Submitted to the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*. UN document CCW/GGE.1/2018/WP.7, 11 April 2018.

Clinciu, Miruna-Adriana, & Helen Hastie. 2019. ‘A Survey of Explainable AI Terminology.’ In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, 8–13. Stroudsburg, Pa.: Association for Computational Linguistics. doi:10.18653/v1/W19-8403

Collopy, Paul, Valerie Sitterle & Jenifer Petrillo. 2020. ‘Validation Testing of Autonomous Learning Systems.’ *Insight* (23) 1: 48–51. As of 27 July 2020: https://onlinelibrary.wiley.com/doi/pdf/10.1002/inst.12285

Comiter, Marcus. 2019. *Attacking Artificial Intelligence: AI’s Security Vulnerability and What Policymakers Can Do About It*. Cambridge, Mass.: Belfer Center for Science and International Affairs, Harvard Kennedy School. As of 12 July 2020: https://www.belfercenter.org/publication/AttackingAI

Congressional Research Service (CRS). 2019. *Artificial Intelligence and National Security*. Washington, DC.

Convention on Certain Conventional Weapons (CCW). 2019. *Final Report*. UN document CCW/MSP/2019/9, 13 December 2019.

Crootof, Rebecca. 2016. ‘War Torts: Accountability for Autonomous Weapons.’ *University of Pennsylvania Law Review* 164 (6): 1347–402.

Cummings, M.L. 2012. ‘Automation Bias in Intelligent Time Critical Decision Support Systems,’ paper, *American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference*, Chicago, 20–22 September 2020. doi:10.2514/6.2004-6313

Daiki, Yokoyama. 2020. ‘Human–Machine Interaction and Human Control: From Engineering to IHL,’ briefing, *Rio Seminar on Autonomous Weapons*, Rio de Janeiro, 20 February 2020.

Danks, David. 2020. ‘How Adversarial Attacks Could Destabilize Military AI Systems.’ IEEE Spectrum, 26 February, 16.51 GMT. As of 4 June 2020: https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/adversarial-attacks-and-ai-systems

Danks, David, & Alex John London. 2017. ‘Algorithmic Bias in Autonomous Systems.’ In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence AI and autonomy track*, 4691–97. doi:10.24963/ijcai.2017/654

Davison, Neil. 2017. ‘A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law.’ *UNODA Occasional Papers* No. 30. New York: United Nations. doi:10.18356/6fce2bae-en

Defense Advanced Research Projects Agency (DARPA). 2016. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*. Arlington, Va.

Defense Innovation Board. 2019. *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense, Supporting Document*. As of 8 September 2020: https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF

Devitt, S. Kate. 2018. 'Trustworthiness of Autonomous Systems.' In *Foundations of Trusted Autonomy (Studies in Systems, Decision and Control)*, Vol. 117, edited by Hussein A. Abbass, Jason Scholz & Darryn J. Reid, 161–84. Cham: Springer.

Dickson, Ben. 2020. 'The advantages of self-explainable AI over interpretable AI.' The Next Web, 19 June. As of 8 September 2020: https://thenextweb.com/neural/2020/06/19/the-advantages-of-self-explainable-ai-over-interpretable-ai

Dietvorst, Berkeley J., Joseph P. Simmons & Cade Massey. 2016. 'Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them.' *Management Science* 64 (3): 983–1476. doi:10.1287/mnsc.2016.2643

Diop, Mouhamadou-Lamine. 2018. 'Explainable AI: The Data Scientists' New Challenge.' Towards Data Science, 14 June. As of 16 June 2020: https://towardsdatascience.com/explainable-ai-the-data-scientists-new-challenge-f7cac935a5b4

Dix, Alan, Janet Finlay, Gregory Abowd & Russell Beale. 2003. *Human Computer Interaction*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Docherty, Bonnie. 2020. 'The Need for and Elements of a New Treaty on Fully Autonomous Weapons.' Human Rights Watch, 1 June. As of 6 June 2020: https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons

Doshi-Velez, Finale, & Been Kim. 2017. 'Towards a Rigorous Science of Interpretable Machine Learning.' arXiv:1702.08608, 2 March. As of 27 July 2020: https://arxiv.org/pdf/1702.08608.pdf

Economist, The. 2020. 'The Potential and the Pitfalls of Medical AI.' 11 June. As of 23 June 2020: https://www.economist.com/technology-quarterly/2020/06/11/the-potential-and-the-pitfalls-of-medical-ai

Ekelhof, Merel, & Giacomo Persi Paoli. 2020a. *Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems*. Geneva: United Nations Institute for Disarmament Research.

———. 2020b. *The Human Element in Decisions about the Use of Force*. Geneva: United Nations Institute for Disarmament Research.

Ethical AI Institute. 2020. 'The AI-RFX Procurement Framework.' Ethical AI Institute. As of 3 July: https://ethical.institute/rfx.html

Feldman, Philip, Aaron Dant & Aaron Massey. 2019. 'Integrating Artificial Intelligence into Weapon Systems.' arXiv:1905.03899v1, 10 May. As of 12 June 2020: https://arxiv.org/pdf/1905.03899.pdf

Ferrini, Mattia. 2019. 'Shall We Build Transparent Models Right Away?' Towards Data Science, 20 August. As of 18 June 2020: https://towardsdatascience.com/shall-we-build-transparent-models-right-away-196db0eeba6c

Fountaine, Tim, Brian McCarthy & Tamim Saleh. 2019. 'Building the AI-Powered Organization.' *Harvard Business Review*, July–August. As of 29 June 2020: https://hbr.org/2019/07/building-the-ai-powered-organization

Fumo, David. 2017. 'Types of Machine Learning Algorithms You Should Know.' Towards Data Science, 15 June. As of 12 June 2020: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

Galyardt, April. 2018. 'Explainable AI and Human Computer Interaction,' briefing, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pa., 24 October 2018. As of 19 May 2020: https://www.youtube.com/watch?v=LnDku7WK0VA

Garcia, Eugenie V. 2019. 'The Militarization of Artificial Intelligence: A Wake-Up Call for the Global South.' SSRN, 10 September. As of 27 July 2020: https://ssrn.com/abstract=3452323

Gebman, Jean R., Douglas W. McIver & Hyman L. Shulman. 1980. *A New View of Weapon System Reliability and Maintainability*. Santa Monica: RAND Corporation.

Google Developers. 2020. 'Machine Learning Crash Course – Generalization.' Google Developers, 10 February. As of 21 June 2020: https://developers.google.com/machine-learning/crash-course/generalization/video-lecture

Goussac, Netta. 2019. 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting.' Humanitarian Law & Policy, 18 April. As of 13 June 2020: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting

Greene, Robert. 2006. *The 33 Strategies of War.* London: Profile Books.

Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS). 2018. *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. UN document CCW/GGE.1/2018/3, 23 October 2018.
———. 2019. *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, 25 September 2019.

Hagström, Martin. 2019. 'Military Applications of Machine Learning and Autonomous Systems.' In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. 1, *Euro-Atlantic Perspectives*, edited by Vincent Boulanin. Stockholm: Stockholm International Peace Research Institute.

Hall, Brian K. 2017. 'Autonomous Weapons Systems Safety.' *Joint Forces Quarterly* 86: 86–93

Haugh, Brian A., David A. Sparrow & David M. Tate. 2018. *The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems*. Alexandria, Va.: Institute for Defense Analyses.

Hawley, John K. 2017. *Patriot Wars: Automation and the Patriot Air and Missile Defense System*. Washington, DC: Center for a New American Security.

Heaven, Douglas. 2019. 'Why Deep-Learning AIs Are So Easy to Fool.' *Nature*, 9 October. As of 25 June 2020: https://www.nature.com/articles/d41586-019-03013-5

Heaven, Will Douglas. 2020. 'Why Asking an AI to Explain Itself Can Make Things Worse.' *MIT Technology Review,* 29 January. As of 6 May 2020: https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse

Herman, Bernease. 2017. 'The Promise and Peril of Human Evaluation for Model Interpretability.' arXiv:1711.07414, 20 November. As of 8 September 2020: https://arxiv.org/pdf/1711.07414v1.pdf

Hoffman, Robert R. 2017. 'A Taxonomy of Emergent Trusting in the Human–Machine Relationship.' In *Cognitive Systems Engineering: The Future for a Changing World*, edited by Philip J. Smith & Robert R. Hoffman. Boca Raton, Fla.: Taylor & Francis.

Hoffman, Robert R., Shane T. Mueller, Gary Klein & Jordan Litman. 2019. 'Metrics for Explainable AI: Challenges and Prospects.' arXiv:1812.04608v2, 1 February. As of 8 September 2020:

Holland Michel, Arthur. 2020. 'The Killer Algorithms Nobody's Talking About.' *Foreign Policy*, 20 January, 6.09 a.m. As of 3 June 2020: https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about

Huh Wong, Yuna, John Yurchak, Robert W. Button, Aaron Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris & Sebastian Joon Bae. 2020. *Deterrence in the Age of Thinking Machines*. Santa Monica: RAND Corporation.

Human Rights Watch (HRW) & International Human Rights Clinic (IHRC). 2015. *Mind the Gap: The Lack of Accountability for Killer Robots*. New York: HRW.

Ilachinski, Andrew. 2017. *AI, Robots, and Swarms Issues, Questions, and Recommended Studies*. Washington, DC: Center for Naval Analyses.

Ingrand, Félix. 2019. 'Recent Trends in Formal Validation and Verification of Autonomous Robots Software,' paper, *Third IEEE International Conference on Robotic Computing,* Naples, 25–27 February 2019. doi:10.1109/IRC.2019.00059

Institute of Electrical and Electronics Engineers (IEEE). 2017a. *IEEE Standard for System, Software, and Hardware Verification and Validation*. IEEE Std 1012-2016, 29 September. New York. doi:10.1109/IEEESTD.2017.8055462.

———. 2017b. 'Reframing Autonomous Weapons Systems.' As of 13 June 2020: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_reframing_autonomous_weapons_v2.pdf

International Committee of the Red Cross (ICRC). 2014. *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. Geneva.

———. 2018. *The Element of Human Control*. Submitted to the Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 20 November. As of 8 September 2020: https://www.unog.ch/80256EDD006B8954/(httpAssets)/810B2543E1B5283BC125834A005EF8E3/$file/CCW_MSP_2018_WP3.pdf

———. 2019a. 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach.' 6 June. As of 8 September 2020: https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach

———. 2019b. *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*. Geneva.
International Organization for Standardization (ISO). 2017. *Artificial Intelligence*. ISO/IEC JTC 1/SC 42. As of 12 June 2020: https://www.iso.org/committee/6794475.html

Jobin, Anna, Marcello Ienca & Effy Vayena. 2019. 'The Global Landscape of AI Ethics Guidelines.' *Nature Machine Intelligence* 1: 389–99. doi:10.1038/s42256-019-0088-2

Joe, Jeffrey C., John O'Hara, Heather D. Medema & Johanna H. Oxstrand. 2014. 'Identifying Requirements for Effective Human Automation Teamwork.' Idaho National Laboratory, June. As of 27 July 2020: https://inldigitallibrary.inl.gov/sites/sti/sti/6101795.pdf

Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach & Jennifer Wortman Vaughan. 2020. 'Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning,' paper 92, *CHI 2020*, 25–30 April 2020.

Kiernan, Kristine M. 2015. 'Human Factors Considerations in Autonomous Lethal Unmanned Aerial Systems,' paper, *Aviation, Aeronautics, and Aerospace International Research Conference*, Phoenix, Ariz., 15–18 January 2015. As of 8 September 2020: https://commons.erau.edu/aircon/2015/Friday/22

Koopman, Philip, & Michael Wagner. 2016. 'Challenges in Autonomous Vehicle Testing and Validation.' *SAE International Journal of Transportation Safety* 4 (1): 15–24. doi:10.4271/2016-01-0128

Kostopoulos, Lydia. 2018. *The Role of Data in Algorithmic Decision-Making: A Primer*. Geneva: United Nations Institute for Disarmament Research.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike & Shane Legg. 2020. 'Specification Gaming: The Flip Side of AI Ingenuity.' DeepMind, 21 April. As of 24 August 2020: https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity

Kuang, Cliff. 2017. 'Can A.I. Be Taught to Explain Itself?' *New York Times*, 21 November. As of 26 May 2020: https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Lage, Isaac, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman & Finale Doshi-Velez. 2019. 'An Evaluation of the Human-Interpretability of Explanation,' arXiv:1902.00006v2, 28 August. As of 8 September 2020: https://arxiv.org/pdf/1902.00006.pdf

Lakkaraju, Himabindu, & Osbert Bastani. 2020. '"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations.' In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, 79–85. New York: Association for Computing Machinery. doi:10.1145/3375627.3375833

Lawand, Kathleen. 2020. 'International Humanitarian Law (IHL) and "LAWS": Is There a Need for a New Protocol?', briefing, *Rio Seminar on Autonomous Weapons*, Rio de Janeiro, 20 February 2020.

LeCun, Yann (@ylecun). 2020. 'We often hear that AI systems must provide explanations and establish causal relationships, particularly for life-critical applications. Yes, that can be useful. Or at least reassuring....1/n.' Eight-tweet thread. Twitter, 5 February, 1:23 p.m. As of 17 June 2020: https://twitter.com/ylecun/status/1225122824039342081

Leslie, David. 2019. *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. London: The Alan Turing Institute. doi:10.5281/zenodo.3240529

Lewis, Lawrence. 2018. 'AI and Autonomy in War: Understanding and Mitigating Risks.' Center for Naval Analyses. As of 8 September 2020: https://www.cna.org/CNA_files/PDF/Understanding-Risks.pdf

Lewis, Michael, Katia Sycara & Philip Walker. 2018. 'The Role of Trust in Human-Robot Interaction.' In *Foundations of Trusted Autonomy (Studies in Systems, Decision and Control)*. Vol. 117, edited by Hussein A. Abbass, Jason Scholz & Darryn J. Reid, 135–59. Cham: Springer

Lipton, Zachary C. 2016. 'The Mythos of Model Interpretability.' arXiv:1606.03490, 10 June. As of 11 June 2020: https://arxiv.org/pdf/1606.03490.pdf

Mane, Shraddha. 2020. 'Deep Neural Networks Are Easily Fooled – Here's How Explainable AI Can Help.' Enterprise AI, 5 June. As of 21 June 2020: https://www.enterpriseai.news/2020/06/05/deep-neural-networks-are-easily-fooled-heres-how-explainable-ai-can-help

Margulies, Peter. 2016. 'Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts.' In *Research Handbook on Remote Warfare*, edited by Jens David Ohlin. Cheltenham: Edward Elgar.

Martin, Bradley, Danielle C. Tarraf, Thomas C. Whitmore, Jacob DeWeese, Cedric Kenney, Jon Schmid & Paul DeLuca. 2019. *Advancing Autonomous Systems: An Analysis of Current and Future Technology for Unmanned Maritime Vehicles*. Santa Monica: RAND Corporation.

Metx, Cade. 2016. 'How Google's AI Viewed the Move No Human Could Understand.' *Wired*, 14 March. As of 13 June 2020: https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand

Miller, Tim. 2019. 'Explanation in Artificial Intelligence: Insights from the Social Sciences.' *Artificial Intelligence* 267: 1–38. doi:10.1016/j.artint.2018.07.007

Mittelstadt, Brent, Chris Russell & Sandra Wachter. 2019. 'Explaining Explanations in AI.' In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 279–88. New York: Association for Computing Machinery. doi:10.1145/3287560.3287574

Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima & Derek Grossman. 2020. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. Santa Monica: RAND Corporation.

Mueller, Shane T., Robert R. Hoffman & Gary Klein. 2019. 'Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI.' Arlington, Va.: Defense Advanced Research Projects Agency. As of 23 May 2020: https://arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf

Musco Eklund, Amanda. 2020. 'Meaningful Human Control of Autonomous Weapon Systems.' FCAS Forum, February. As of 8 September 2020: http://www.fcas-forum.eu/publications/Meaningful-Human-Control-of-Autonomous-Weapon-Systems-Eklund.pdf

National Institute of Standards and Technology (NIST). 2019. 'FAQs about NIST's Role in Planning Federal Engagement in AI Standards Development.' NIST, 12 August. As of 17 June 2020: https://www.nist.gov/topics/artificial-intelligence/ai-standards/faqs-about-nist-and-ai-standards

National Security Commission on Artificial Intelligence (NSCAI). 2020. *First Quarter Recommendations*. Arlington, Va.

Okamura, Kazuo, & Seiji Yamada. 2020. 'Adaptive Trust Calibration for Human-AI Collaboration.' *PLoS ONE* 15 (2). doi:10.1371/journal.pone.0229132

O'Sullivan, Liz. 2019. 'The Killer Robots Are Coming. Here's Our Best Shot at Stopping Them.' NBC Think, 2 July. As of 22 June 2020: https://www.nbcnews.com/think/opinion/killer-robots-are-coming-here-s-our-best-shot-stopping-ncna1025436

Petsiuk, Vitali, Abir Das & Kate Saenko. 2018. 'RISE: Randomized Input Sampling for Explanation of Black-Box Models.' arXiv:1806.07421v3, 25 September. As of 8 September 2020: https://arxiv.org/pdf/1806.07421.pdf

Phillips, P. Jonathon, Carina A. Hahn, Peter C. Fontana, David A. Broniatowski & Mark A. Przybocki. 2020. 'Four Principles of Explainable Artificial 3 Intelligence.' National Institute of Standards and Technology, August. As of 8 September 2020: https://doi.org/10.6028/NIST.IR.8312-draft

Pontin, Jason. 2018. 'Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning.' *Wired*, 2 February. As of 1 June 2020: https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning

Poursabzi-Sangdeh, Forough, Dan Goldstein, Jake Hofman, Jennifer Wortman Vaughan & Hanna Wallach. 2018. 'Manipulating and Measuring Model Interpretability.' Microsoft Research, February. As of 8 September 2020: https://www.microsoft.com/en-us/research/publication/manipulating-and-measuring-model-interpretability

Quiñonero-Candela, J., M. Sugiyama, A. Schwaighofer & N.D. Lawrence. 2009. *Dataset Shift in Machine Learning*. Cambridge: MIT Press.

Rai, Arun. 2020. 'Explainable AI: From Black Box to Glass Box.' *Journal of the Academy of Marketing Science* 48: 137–41. doi:10.1007/s11747-019-00710-5

Ribeiro, Marco, Sameer Singh & Carlos Guestrin. 2016. '"Why Should I Trust You?": Explaining the Predictions of Any Classifier.' In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. Stroudsburg, Pa.: Association for Computational Linguistics. doi:10.18653/v1/N16-3020

Ribera, Mireia, & Agata Lapedriza. 2019. 'Can We Do Better Explanations? A Proposal of User-Centered Explainable AI,' paper, *ACM IUI 2019 Workshops*, Los Angeles, 20 March 2016.

Rickli, Jean-Marc. 2019. 'The Destabilizing Prospect of Artificial Intelligence for Nuclear Strategy, Deterrence, and Stability.' In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. 1, *Euro-Atlantic Perspectives*, edited by Vincent Boulanin. Stockholm: Stockholm International Peace Research Institute.

Roff, Heather. 2020. 'AI Deception: When Your Artificial Intelligence Learns to Lie.' IEEE Spectrum, 24 February. As of 15 June 2020: https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/ai-deception-when-your-ai-learns-to-lie

Roff, Heather M., & David Danks. 2018. '"Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems.' *Journal of Military Ethics* 17 (1): 2–20. doi:10.1080/15027570.2018.1481907

Roff, Heather M., & Richard Moyes. 2016. 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons.' Article 36. As of 18 June 2020: www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf

Royal Society, The. 2019. *Explainable AI: The Basics*. London. As of 18 June 2020: https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

Rudin, Cynthia. 2019. 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.' *Nature Machine Intelligence* 1: 206–15. doi:10.1038/s42256-019-0048-x

Scharre, Paul. 2020. 'The Militarization of Artificial Intelligence.' In 'Policy Roundtable: Artificial Intelligence and International Security'. *Texas National Security Review*, 2 June. As of 23 May 2020: https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security

Schmelzer, Ron. 2019. 'Understanding Explainable AI.' *Forbes*, 23 July, 7.12 a.m. EDT. As of 29 May 2020: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai

———. 2020. 'Towards a More Transparent AI.' *Forbes,* 23 May, 1.28 p.m. EDT. As of May 26: https://www.forbes.com/sites/cognitiveworld/2020/05/23/towards-a-more-transparent-ai

Selbst, Andrew D., & Julia Powles. 2017. 'Meaningful Information and the Right to Explanation.' *International Data Privacy Law* 7 (4): 233–42. doi:10.1093/idpl/ipx022

Sessions, Valerie, & Marco Valtorta. 2006. 'The Effects of Data Quality on Machine Learning Algorithms,' paper, *11th International Conference on Information Quality*, Cambridge, Mass., 10–12 November 2006.

Simonite, Tom. 2018. 'Google's AI Guru Wants Computers to Think More Like Brains.' *Wired,* 12 December. As of 17 June 2020: https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains

Slayton, Rebecca. 2020. 'The Promise and Risks of Artificial Intelligence: A Brief History.' In 'Policy Roundtable: Artificial Intelligence and International Security'. *Texas National Security Review*, 2 June. As of 23 May 2020: https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security

Snoek, Jasper, & Zachary Nado. 2020. 'Can You Trust Your Model's Uncertainty?' Google AI Blog, 15 January. As of 21 May 2020: https://ai.googleblog.com/2020/01/can-you-trust-your-models-uncertainty.html

Sokol, Kacper, & Peter Flach. 2020. 'One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency.' *Künstliche Intelligenz* 34: 235–50. doi:10.1007/s13218-020-00637-y

Sparrow, Robert. 2015. 'Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender.' *International Law Studies* 91: 699–728.

Stewart, Matthew. 2019. 'Understanding Dataset Shift.' Towards Data Science, 11 December. As of 15 August 2020: https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766

Stubbs, Kristen, Pamela J. Hinds & David Wettergreen. 2007. 'Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *IEEE Intelligent Systems* 22 (2): 42–50. doi:10.1109/MIS.2007.21

Tjoa, Erico, & Cuntai Guan. 2020. 'A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI.' arXiv:1907.07374v4, 7 June. As of 8 September 2020: https://arxiv.org/pdf/1907.07374v4.pdf

Trumbull, Charles P., IV. 2020. 'Autonomous Weapons: How Existing Law Can Regulate Future Weapons.' *Emory International Law Review* 34 (2). doi:10.2139/ssrn.3440981

United Nations Institute for Disarmament Research (UNIDIR). 2014. *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*. Geneva.

———. 2016. *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*. Geneva.

———. 2017. *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches*. Geneva.

US Department of Defense. 2017. *Directive 3000.09: Autonomy in Weapon Systems*. Washington, DC. As of 13 July 2020: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf

Van den Bosch, Karel, & Adelbert Bronkhorst. 2018. 'Human-AI Cooperation to Benefit Military Decision Making.' Brussels: Science & Technology Organization, NATO.

Wang, Ning, David V. Pynadath & Susan G. Hill. 2016. 'Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations.' In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 109–16. New York: IEEE. doi:10.1109/HRI.2016.7451741

Williams, Andrew P. 2015. 'Defining Autonomy in Systems: Challenges and Solutions.' In *Autonomous Systems: Issues for Defence Policymakers*, edited by Andrew P. Williams & Paul D. Scharre. Brussels: Headquarters Supreme Allied Commander Transformation, NATO.

Wu, Weibin, Hui Xu, Sanqiang Zhong, Michael R. Lyu & Irwin King. 2019. 'Deep Validation: Toward Detecting Real-World Corner Cases for Deep Neural Networks.' In *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 125–37. New York: IEEE. doi:10.1109/DSN.2019.00026

Yampolskiy, Roman V. 2019. 'Unpredictability of AI.' arXiv:1905.13053, 29 May. As of 8 September 2020: https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf

———. 2020. 'Unexplainability and Incomprehensibility of Artificial Intelligence.' Mind Matters,10 February. As of 5 August 2020: https://mindmatters.ai/2020/02/unexplainability-and-incomprehensibility-of-ai

Zhang, Quan-shi, & Song-chun Zhu. 2018. 'Visual Interpretability for Deep Learning: A Survey.' *Frontiers of Information Technology & Electronic Engineering* 19: 27–39. doi:10.1631/FITEE.1700808

THE BLACK BOX, UNLOCKED

PREDICTABILITY AND UNDERSTANDABILITY IN MILITARY AI

Predictability and understandability are widely held to be vital characteristics of artificially intelligent systems. Put simply: AI should do what we expect it to do, and it must do so for intelligible reasons. This consideration stands at the heart of the ongoing discussion about lethal autonomous weapon systems and other forms of military AI. But what does it mean for an intelligent system to be "predictable" and "understandable" (or, conversely, unpredictable and unintelligible)? What is the role of predictability and understandability in the development, use, and assessment of autonomous weapons and other forms of military AI? What is the appropriate level of predictability and understandability for AI weapons in any given instance of use? And how can these thresholds be assured?

This study provides a clear, comprehensive introduction to these questions, and proposes a range of avenues for action by which they may be addressed.

40
UNIDIR