DATA ISSUES AND MILITARY AUTONOMOUS SYSTEMS

ARTHUR HOLLAND MICHEL

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

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. Rebecca Crootof, Dr. S. Kate Devitt and Dr. Maria Vanina Martinez, 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.

ABOUT UNIDIR

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

NOTE

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

CITATION

Holland Michel, Arthur. 2021. *Known Unknowns: Data Issues and Military Autonomous Systems*. Geneva: United Nations Institute for Disarmament Research. doi.org/10.37559/SecTec/21/AI1

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.

KEY TERMS

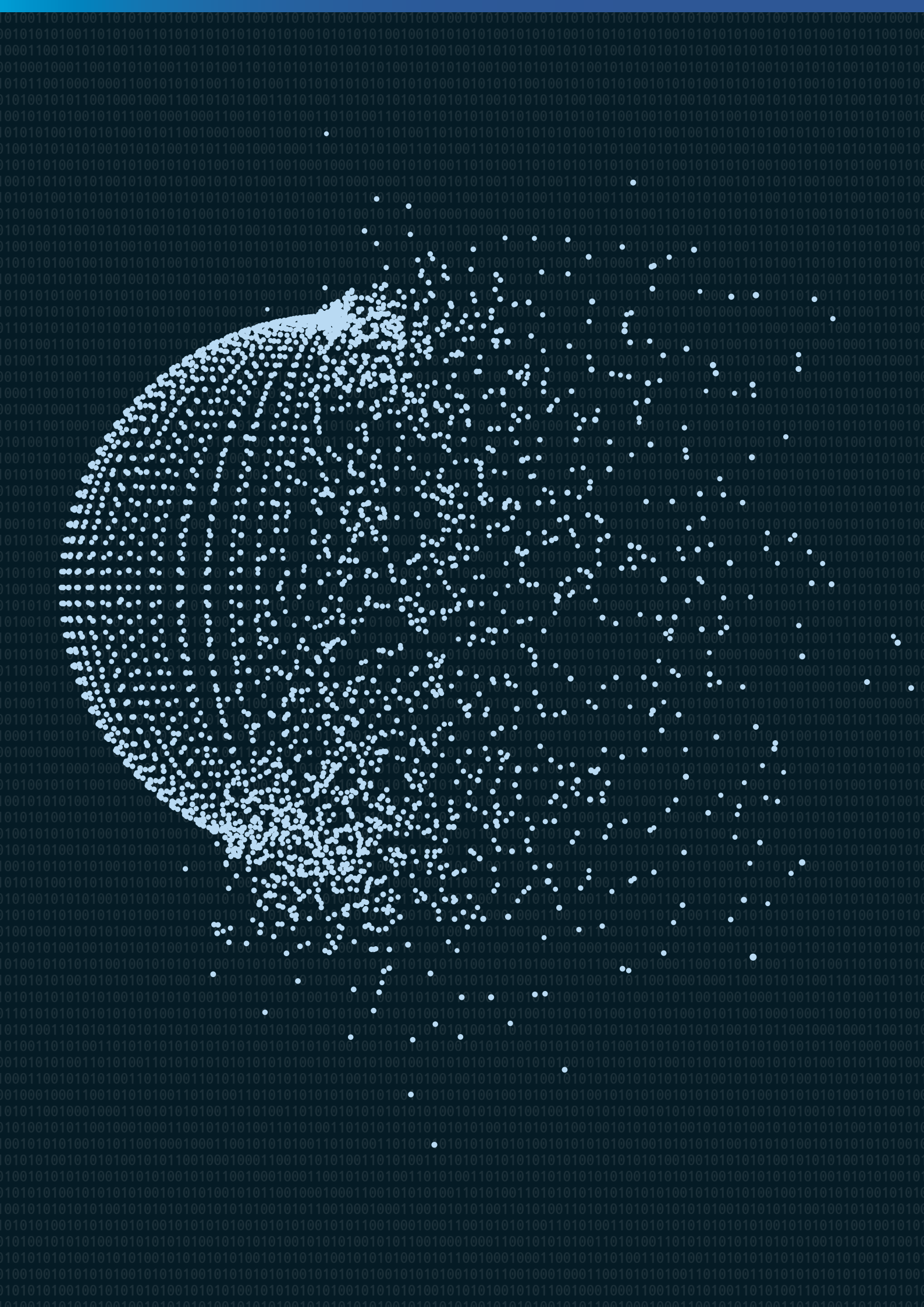
Data issues	errors or problems in the data that an autonomous system receives during an operation.
Vulnerability	a flaw in an autonomous system that causes it to fail when it encounters a certain data issue.
Failure	occurs when an autonomous system does not exhibit the desired behaviour.
Known unknowns	vulnerabilities that exist in an autonomous system but were not specifically identified in advance.
Brittleness	an autonomous system's tendency to fail when encountering inputs for which it was not specifically designed or tested.
Robustness	an autonomous system's ability to perform as desired when receiving inputs for which it was not designed or tested.
Uncontrolled environments	environments where the precise conditions to which the system is subjected cannot be tightly managed by human operators — in contrast with, for instance, a laboratory setting where conditions (such as physical features of the space, temperature, lighting, presence of other systems) are tightly controlled.

"THE INTELLIGENCE IS IN THE DATA, NOT THE ALGORITHM."

– HAUGH ET AL. (2018)

All autonomous systems run on data. Indeed, one way to think of "autonomy" is as a process by which machines respond to data inputs from their environment with a corresponding output or action, without any human direction. If there are issues with these data inputs, autonomous systems can exhibit suboptimal performance or fail.

In the real world, data are never perfect. More importantly, they are imperfect in complex and unpredictable ways. Autonomous systems encountering data issues may likewise fail in a complex and unpredictable manner. As a result, autonomous system failures arising from data issues could be both inevitable *and* impossible to anticipate; these are the "known unknowns".



INTRODUCTION

The vagaries of data are central to the ongoing discussion among policymakers about the harms that could arise from the use of autonomous weapon systems (AWS) and other forms of military artificial intelligence (AI). To be sure, it has been posited that autonomous systems could potentially exhibit better performance in certain tasks than traditional means or methods of warfare, leading to a reduction in some kinds of unintended harm. But because data issues can result in autonomous system failures that could *increase* harm,[1] such issues are relevant to deliberations over whether such technologies could "perform tasks as expected or be capable of being used in accordance with [international humanitarian law]".[2]

Data issues and the failures they cause are also entwined with any notion of direct or indirect human control or judgment, a fundamental principle for the use of all lethal autonomous weapons. In any military operation, the possibility of system failures must factor into the human decision on whether and how to use such weapons.[3] Furthermore, the ability of autonomous systems to adhere to the constraints or guardrails that their human operators set for them[4] will, itself, depend on reliable data. More fundamentally, although humans could exercise forms of control over autonomous systems throughout all the stages of their development and use,[5] one can never fully control the real-world data that these systems depend on to function properly and reliably. Indeed, though it is difficult, as many have pointed out, to anticipate exactly what form autonomous weapons will take in future, these systems will always contend with data that are problematic and unpredictable.

As such, any potential future policy related to human control,the application of international law, or the "operationalization" of the GGE's guiding principles, will likely have to account for data issues. But by and large, discussions in this policy domain still lag behind the science of the matter.[6]

This report is provided for the policy community to advance the state of understanding of these issues and their implications in discussions on autonomous weapons and military AI. It is based on interviews with technical, military and legal subject matter experts, as well as an extensive review of the relevant academic and policy literature.

The report finds that in order to perform as desired, autonomous systems must collect data that are complete, relevant, accurate and of high quality; most importantly, these data must not differ from the data for which a system was developed and tested. But conflict environments are harsh, dynamic and adversarial, and there will always be more variability in the real-world data of the battlefield than in the limited sample of data on which autonomous systems are built and verified. Because they are complex systems, autonomous weapons encountering such unavoidable data issues could likewise fail in a complex and unpredictable manner. As such, all autonomous systems will be prone to inevitable accidents which cannot be foreseen. We know that the potential for such accidents exist either now or will emerge in the future, but we cannot characterize or specifically anticipate them. One might call such issues "known unknowns".

1 Scharre (2016, 5).

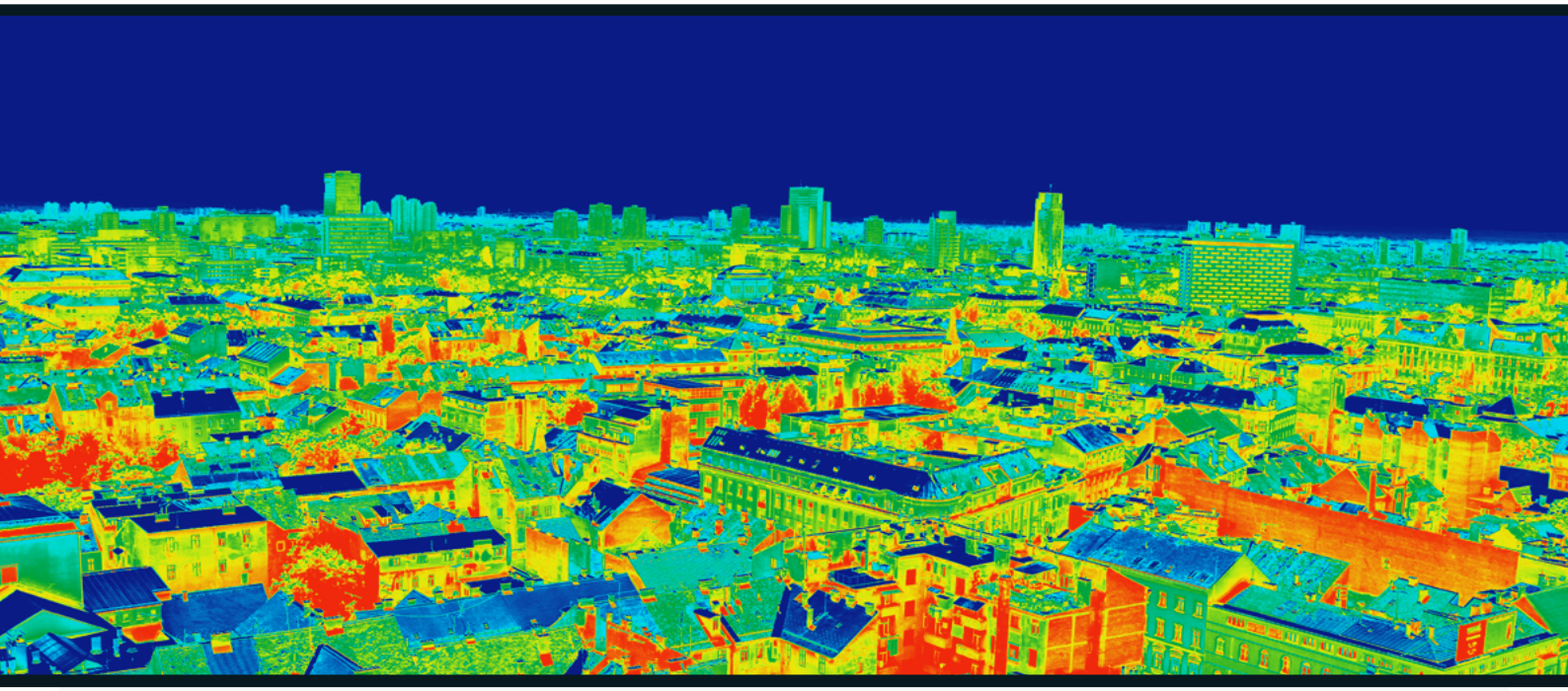
2 GGE on LAWS (2019, annex IV).

3 GGE on LAWS (2020, 4) states that "an effective response to the risks posed by autonomous weapon systems, thus, may require consideration of *what* 'quality and extent' of human control/involvement/judgment is necessary".

4 GGE on LAWS (2020, 4) notes that "there is an emerging consensus [among States]…that limits on emerging technologies in the area of [lethal autonomous weapons systems] are required in order to ensure compliance with [international humanitarian law] and other applicable law".

5 GGE on LAWS (2018, 15).

6 While the relevant policy forums have visited the issue of data repeatedly (particularly with respect to bias and adversarial hacking), data issues take a much broader variety of forms and have a broader set of implications than those areas of focus. GGE on LAWS (2019, 3, 5); GGE on LAWS (2020, 7).



These technical realities have potentially profound implications. International law requires States employing autonomous weapons to anticipate and respond to data issues that could cause unintended harm. A variety of complex interrelated factors determine the degree to which States could address such issues. More fundamentally, the ability *and thus responsibility* to mitigate or account for the potential harms arising from data issues hinges on whether these issues are known *or* unknown. The fact that some autonomous system vulnerabilities are "known unknowns" could create ambiguity as to responsibility for unintended harm resulting from such issues.

Data issues could therefore pose a novel challenge to the responsible, legal and human-centric employment of autonomous military systems. In Chapter 5, the report examines the variety of approaches that have been proposed to address this challenge. In the Conclusion, the report recommends five avenues for action to help bolster these and other potential measures and efforts. Specifically, the report calls for collaborative, science-based action relating to:

- Legal reviews
- Classification of autonmous systems accidents
- Knowledge-sharing between States
- Understanding and modelling of the effects of adversarial countermeasures against autonomous systems
- Consideration of autonomous weapon systems as system-of-systems technology

1. COMMON DATA ISSUES

Autonomous systems rely on the data they collect in order to navigate,[7] interpret, respond to and manipulate their environment. Data are also the medium by which they receive human control, monitor their own internal system state and health,[8] and determine their progress towards their goal.[9] Issues can manifest themselves in any of the data that underpin these essential functions. These issues either arise naturally or as a result of intentional adversarial measures.

Broadly speaking, these issues can be categorized as incomplete data, low-quality data, incorrect or false data, and discrepant data (data that differ from the data the system was designed for). This chapter provides a brief overview of each of these types of data issue.

While autonomous systems can, of course, be expected to handle some kinds of data issues, all autonomous systems can be expected to have vulnerabilities to specific or systemic data issues at some point over the course of their employment, even when these systems are used exactly as intended.[10] This is especially true of issues that have not been considered or covered during that system's development or testing.

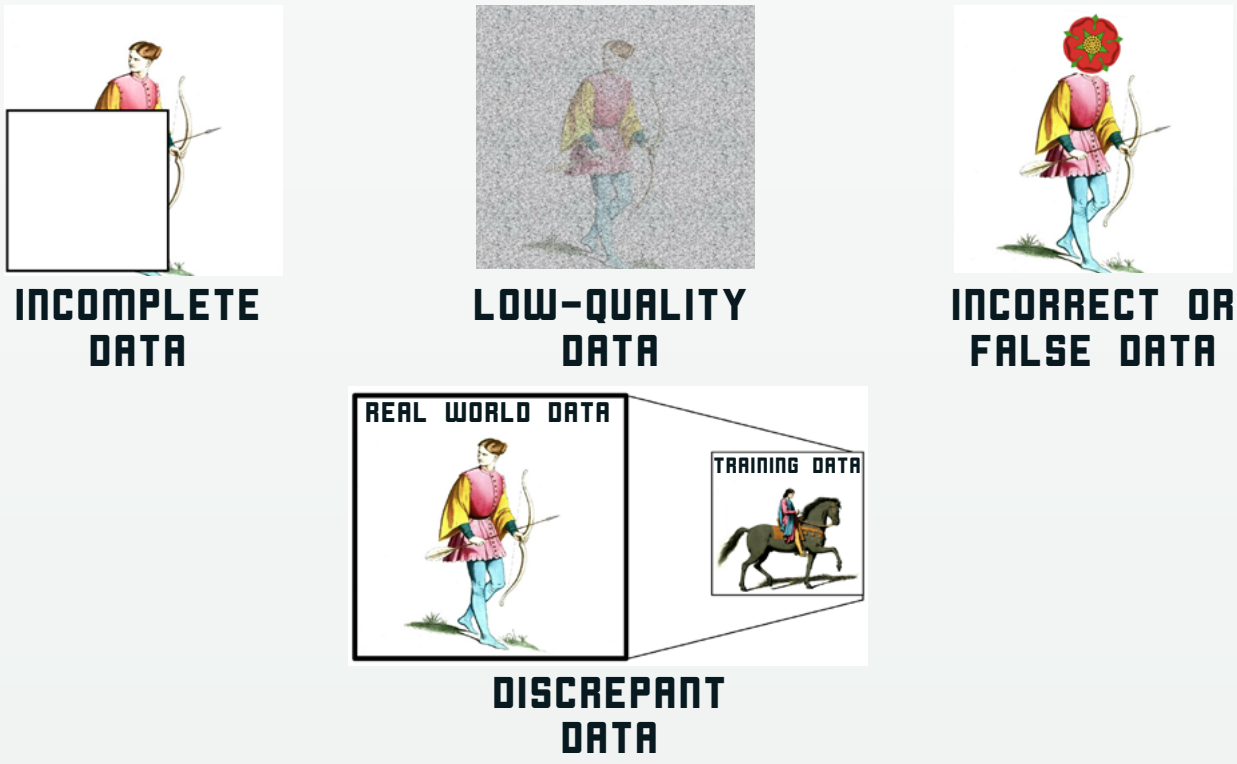


FIGURE 1. *Types of common data issues.*

7 Hagström (2019, 33–34).

8 Atyabi et al. (2020); Ilachinski (2017, 171); Naval Studies Board (2005, 47).

9 Amodei et al. (2016).

10 Interview with Davide Scaramuzza, 2 October 2020; interview with anonymous expert, 18 September 2020; Cummings (2020a).

1.1 INCOMPLETE DATA

Data are "incomplete" when information that is necessary for an autonomous system to take an appropriate action are not present in that data,[11] either because the source of the data is blocked or because the autonomous system lacks the sensors necessary to detect the information or to perceive it.[12]

Missing data may cause an autonomous system to misclassify objects and activities or fail to recognize its progress towards a given goal. Even when systems have some capacity to perceive that a data point is missing, they may struggle to infer what might exist in the gaps in their perception.[13]

1.2 LOW-QUALITY DATA

Data can contain errors or ambiguities that result in suboptimal or inappropriate responses by the autonomous system.[14] A data point may not have sufficient resolution or accuracy, meaning that it fails to capture the exact characteristics of the sensed object or phenomenon. Or the quality of the input from a sensor or data stream may be degraded as a result of extraneous data points (known as "noise" or "clutter") that are irrelevant to the relevant information (the "signal"). In some cases, these extraneous data points arise from the sensor itself (consider, for example, the white noise on a bad phone line) or because the environment is cluttered with irrelevant objects surrounding the object of interest.

1.3 INCORRECT OR FALSE DATA

Data can be incorrect or false. Such instances can arise as a result of common faults in the sensors themselves or in the source of the data. A badly calibrated[15] or faulty sensor[16] feeding an autonomous system might generate an incorrect measurement (such as the size, shape or speed of a target), or a human-generated data feed may include errors (incorrect numbers, spelling mistakes, incorrect formatting, etc.).[17] Incorrect or false data may likewise arise from intentional adversarial actions that are intended to "fool" autonomous systems into making an erroneous output (see Section 2.2).

1.4 DISCREPANT DATA

Autonomous systems today are liable to fail when there is inconsistency between the data they are designed and developed for and data they encounter in actual use.[18]

Systems may encounter one-off inputs or unique combinations of inputs that fall outside the total spectrum of possible inputs that the system was designed for; these are known as anomalies,[19] "edge cases" or "corner cases".[20] Other inputs may be discrepant in the sense that they do not fit neatly within the structured categories that human designers code AI systems to recognize or respond to.[21]

In other cases, such discrepancies may be systemic. This happens when a system is deployed in a place or a manner for which it was not designed[22] or when the system's development

11 Interview with Davide Scaramuzza, 2 October 2020; interview with anonymous expert, 15 October 2020; interview with anonymous expert, 21 October 2020; interview with anonymous expert, 6 November 2020.

12 Schwarz (2018).

13 A classic challenge in the development of automated object tracking algorithms – which can be used, say, for tracking an enemy aircraft in a dogfight – is ensuring that the algorithm does not lose track of the object when it momentarily disappears from view. Pan & Hu (2007).

14 For a foundational discussion of data quality, see Wand & Wang (1996).

15 Jain et al. (2019).

16 Per Bagchi et al. (2020), sensor faults arise naturally as a result of bugs, ageing and other unavoidable environmental factors.

17 Cole (2019, 30); Gates & Baker (2019).

18 Amodei et al. (2016, 16); Lohn (2020a); Taori et al. (2020, 1, 8).

19 For a technical, statistics-focused discussion of different types of anomaly, see Chandola et al. (2009, 7–8).

20 One of the reasons that the training data sets used to "teach" machine-learning systems must be very large, and why physical testing must be extensive, is to reduce the likelihood that the system will encounter something it was not trained for (See Annex I). Gershgorn (2017).

21 For example, a system trained to distinguish between trucks and tanks may struggle with an armoured personnel carrier that exhibits characteristics of both categories. Llorens (2020); Schwarz (2018).

22 This phenomenon is known as "transfer context bias". Danks & London (2017, 4694); UNIDIR (2018, 4).

process has failed to accurately and fairly capture key characteristics of the intended environment. (For example, many instances of "biased AI" arise because a particular demographic group was under- or mis-represented in the system's training data.[23])

While autonomous systems could be developed to be more tolerant of issues like incomplete data, low-quality data or false data, they will only be robust to those issues that were specifically accounted for in their development or training. **Any issues that were not covered in a system's development could still cause a failure.**

1.5 DATA ISSUE "AWARENESS"

The data issues described in this chapter result in failures when the autonomous system has not been developed to account for such issues. In these cases, the system will not be "aware" that it is encountering an issue.[24] For example, if an AI system is trained to recognize when an object is partly hidden, it could be coded to revert to a failsafe whenever it encounters such instances of incomplete data. However, in the absence of any such "awareness", an autonomous system will generate what might be described as a "best guess" attempt at a solution.

That is, if an autonomous system encounters an unfamiliar object, it will simply misclassify that object as whatever most resembles it; in practice, it would be more desirable if it were to label the object as an "unclassified object".[25] Recent advances in "hybrid" autonomous systems that incorporate multiple types of AI have shown some potential to address the problem (see Annex II), but this remains an open research field.[26]

23 For a detailed discussion of types of algorithmic bias and their implications for autonomous weapons, see UNIDIR (2018). For a discussion of the ethical implications of AI bias, see Buolamwini & Gebru (2018, 11–12); Grother et al. (2019). Note that not all biases in algorithmic models are "bad". Indeed, all such models must be embedded with certain biases to function properly. See Hellström et al. (2020); Krishnamurthy (2019).

24 Interview with anonymous expert, 6 November 2020.

25 For this notion of "reasoning", see Cummings (2020a, 4–5). This is also why certain adversarial examples would only be effective against automated systems and not humans. For example, in the case of physical visual spoofing (see Figure 4) the "disguises" that have proven effective against computers may be easily detectable by humans. Though other anomalies (such as the addition of subtle noise to large data sets) may be very hard for humans to perceive. See Haugh et al. (2018, 2–3); Lohn (2020b, 11–12).

26 Interview with Maria Vanina Martinez, 2 November 2020.

2. CAUSES OF DATA ISSUES

Compared with the controlled, often digital, environments where AI has proven itself so far – such as social media, finance, insurance, medicine, manufacturing and gameplay – "uncontrolled" conflict environments pose a wide range of challenges to the collection of complete, true, high-quality, non-discrepant data.[27] This is because all such conflict environments are harsh, adversarial, complex and variable.

2.1 HARSH CONDITIONS

In warfare, autonomous systems will be subject to challenging environmental factors that will inhibit the collection of reliable and consistent data.

- Dust, smoke, vibrations, contaminants, kinetic effects and adverse weather can obscure or damage vital sensors and the instruments and subsystems with which they interact.[28]
- Natural wear and tear can degrade sensor inputs.[29]
- Objects or data points of interest may rarely appear in the full view or range of the sensors.[30]
- Camouflage and concealment will often obscure relevant events or objects.

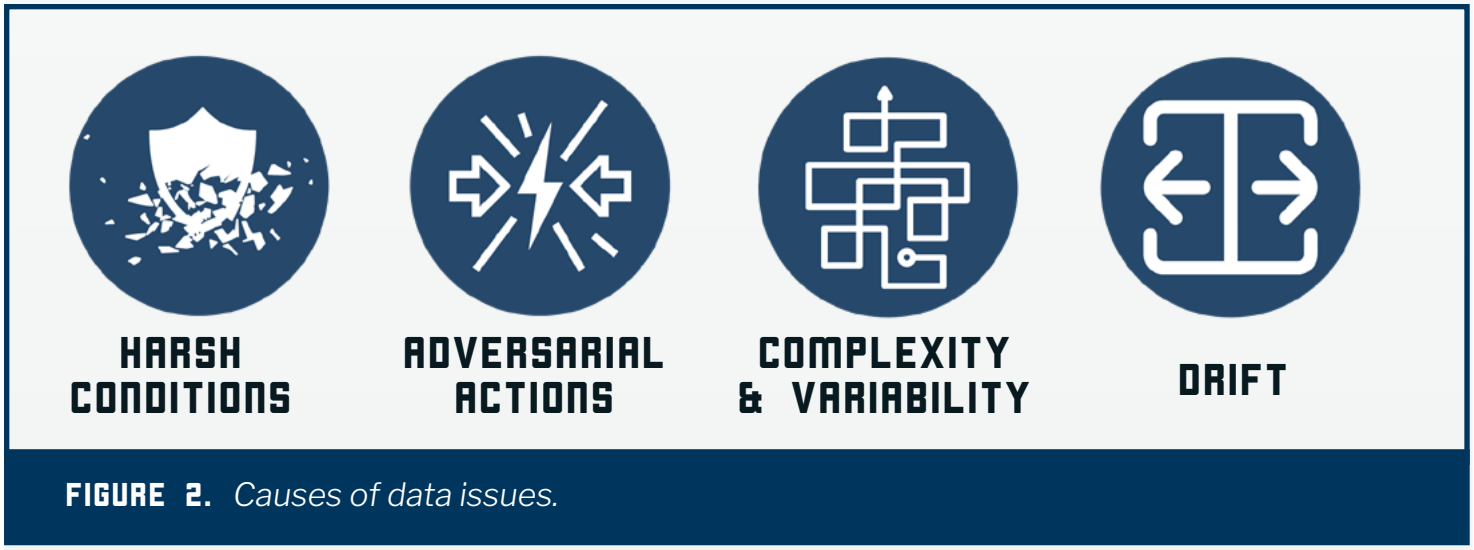


FIGURE 2. *Causes of data issues.*

27 Successful AI deployments in simulated settings – for example, the game-playing AlphaGo/AlphaGo Zero/AlphaZero computers, the GPT-3 language generation model, or the automated pilot that beat a human fighter pilot in a series of virtual dogfights – are not an accurate indicator of the deftness, consistency and reliability with which such systems could excel in military applications. A more accurate yardstick for the progress of autonomous physical systems is self-driving cars, which – despite billions of research and development investment and millions of kilometres of testing – remain a long way from fully scaled deployment and integration. Interview with Davide Scaramuzza, 2 October 2020; interview with Maria Vanina Martinez, 2 November 2020; Cummings (2020c). For more on the challenge of developing robots that can execute complex strategies in complex environments, see Ibarz et al. (2021).

28 Interview with anonymous expert, 6 November 2020; Defence and Security Accelerator (2019); French et al. (2016).

29 Ibarz et al. (2021).

30 Ramisa Ayats et al. (2012, 187).

2.2 ADVERSARIAL ACTION

In conflict settings, autonomous systems will be subjected to targeted countermeasures that will give rise to data issues.

- Belligerents will target autonomous systems with kinetic effects, causing damage that may degrade their sensors or data receivers, or target their sensors specifically.[31]
- Signal jamming would block systems from receiving certain data inputs (especially navigation data) or communications from their operators[32] – a vulnerability for any system that relies on a human in or on the loop.
- Like any computer-enabled system, autonomous systems will also be susceptible to hacking; "spoofing" attacks, for example, would replace an autonomous system's real incoming data feed with a fake feed containing incorrect or false data.[33]
- Belligerents will seek to take advantage of the brittleness of autonomous systems by simply modifying their actions[34] or "poisoning" the data in such a way as to generate discrepant inputs for which the system is not designed.[35]
- Autonomous systems will also be targeted by a class of adversarial action known as "input attacks", which change a sensed object or data source in such a way as to generate a failure.[36]

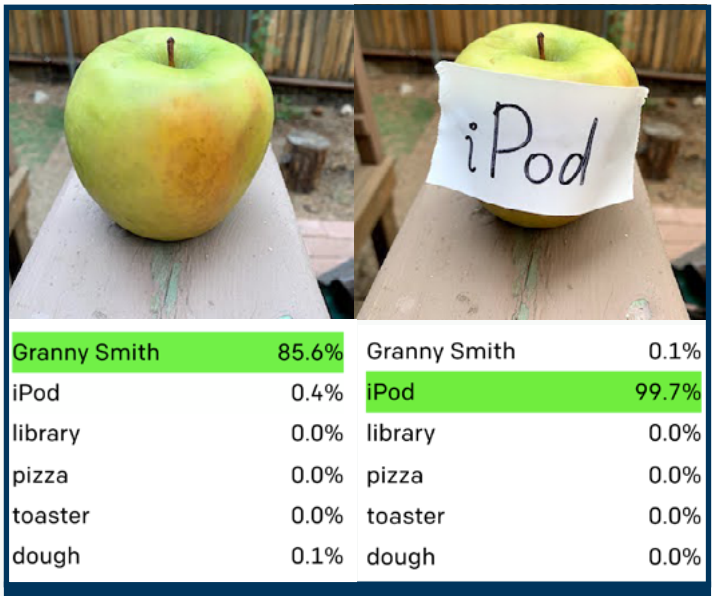


FIGURE 3. *A demonstration by Goh et al. (2021) showing that a sophisticated neural network could be fooled into misclassifying an apple with a handwritten label.*
CREDIT: OPENAI

31 Optical sensors, for example, are susceptible to bright flashes of light from lasers and floodlights, a type of attack known as "dazzling". Birch et al. (2015, 18); Tholl (2018).

32 Holland Michel (2020a).

33 GNSS (Global Navigation Satellite System) data spoofing is already a well-studied and routinely employed countermeasure against systems that rely on GNSS data for navigation. Jafarnia-Jahromi et al. (2012); Kerns et al. (2014).

34 A classic example of this practice is the continuous efforts of email spam creators to devise new formats for their messages that will not be recognizable to AI spam filters. Kantchelian et al. (2013); see also Herpig (2019, 20). Adversarial actors can also use their exposure to an autonomous system to train their own "adversarial models" that can, in turn, generate tailored inputs to flummox that system: see Bagchi et al. (2020).

35 Goldblum et al. (2020).

36 Simple input attacks might seek to confound an autonomous system by disguising a target. A more sophisticated type of attack known as "adversarial examples" or "evasion" involve adding subtle artefacts to an input datum that result in catastrophic interpretation error by the machine. See Athalye et al. (2017); Bhambri et al. (2019); Herpig (2019, 19); Li (2019); Lohn (2020b). Numerous research projects in recent years have demonstrated the effectiveness of such attacks in causing sophisticated machine-learning systems to misclassify objects with a high degree of confidence, in some cases without changing the datum so much that it would appear tainted to a human reviewing the input. Such attacks can either be developed based on a specific detailed understanding of an autonomous system's algorithmic models (this is known as a "white box" attack) or without any understanding of or access to the system's code or training data ("black box" attacks).

2.3 COMPLEXITY AND VARIABILITY

Warfare environments are extremely complex.[37] The more complex an operational environment, the greater the degree and diversity of relevant information a system must take into account[38] to achieve its goals. But the more information the system requires, the higher the likelihood that some of that data will be occluded, corrupted or discrepant. Furthermore, these autonomous machines will be complex systems that rely on seamless interaction between multiple types of tightly coupled[39] sensors, algorithms, actuators and human factors – all of which could, in their own way, be vulnerable to certain kinds of data issues that in isolation would not result in system failure.[40]

More fundamentally, there will always be more potential variability in a deployed military autonomous system's environment than there was in its finite development environment.[41] This means that some issues will only manifest and cause failures once the autonomous system is deployed.[42]

- The data sets used to train and test machine-learning systems can only, at best, capture a statistical approximation of reality (the same way that a sample survey provides an approximate, but never perfect, depiction of an entire population).[43]
- Rule-based coding can only approximate the subtle and changeable dynamics of real, physical conditions.

37 Danzig (2018, 7).

38 Cummings (2017, 4).

39 "Coupling" refers to the interdependency of components: in a tightly coupled system, the behaviours of one component directly affect the components that it interacts with. UNIDIR (2016, 6). Herpig (2019, 31) describes how an attack on a single algorithmic component of a complex system composed of multiple AI elements could have catastrophic cascading effects throughout the system.

40 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Rebecca Crootof, 18 November 2020; Scharre (2016, 13); UNIDIR (2016, 6).

41 For example, large data sets of training images for automated satellite imagery analysis programs cannot include examples of every single object as seen from every single angle at which a satellite might observe the object in real life, and even slight variations in angle between training images and real-life images have been shown to result in a dramatic drop in accuracy. Weir (2018).

42 Danzig (2018, 7).

43 Interview with Maria Vanina Martinez, 2 November 2020; interview with Davide Scaramuzza, 2 October 2020; Ibarz et al. (2021). Lipton (2020) suggests that the notion that a statistical distribution actually even exists in the world is flawed, given that distributions are always changing (see Section 2.4).

› Real-life testing cannot validate systems against all potential issues that a complex autonomous system might encounter in a complex environment, especially against issues that are not known to be issues[44] or the equally varied possibilities for adversarial interference.[45]

2.4 DATA DRIFT

Even if an autonomous system were designed and tested to fully account for the complexity of the real world, environments change[46] in ways that will eventually subject systems to one-off cases or systematic discrepancies that did not previously exist.[47] This phenomenon is known as "data drift".[48] Conflict environments are likely to drift constantly.

› Wartime activities physically change the environment.[49]
› Groups engage in unpredictable behaviour to deceive or surprise the adversary[50] and continually adjust (and sometimes radically overhaul) their tactics and strategies to gain an edge.
› Because drift can happen gradually (or, if adversarial, covertly), it may be difficult to detect.[51]
› Conversely, sudden unanticipated tectonic shifts – for example, the emergence of a wholly novel military tactic for which an autonomous system was not developed or designed – can render whole classes of system ineffective.[52]

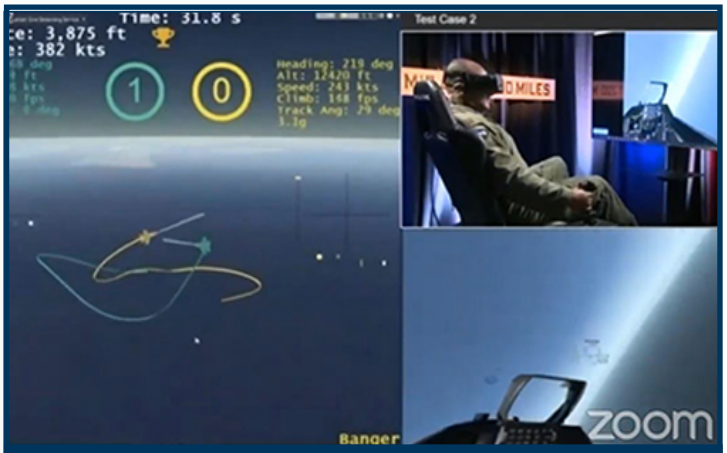


FIGURE 4. *In a set of simulated dogfights between an expert human pilot and an AI "pilot," the machine won 5-0. However, this system was operating in a controlled simulated environment. Real-world uncontrolled environments will prove much more challenging for autonomous systems.*
CREDIT: DEFENSE ADVANCED RESEARCH PROJECTS AGENCY.

44 Interview with Davide Scaramuzza, 2 October 2020; interview with anonymous expert, 18 September 2020; interview with anonymous expert, 21 October 2020.

45 Herpig (2019, 29, 35).

46 This varies in part depending on the nature of the objects or phenomena to which a system is responding. One anonymous expert (interviewed 15 October 2020) described three types of data: static (such as geographical features), highly predictable (such as crowd dynamics or group behaviours) and highly unpredictable (such as objects and signals in a contested battlespace with active denial, deception and subterfuge.)

47 Certain AI models may be particularly sensitive to such changes, particularly if they rely on the presence of a single environmental factor to achieve a desired output or if a new factor that the model does not take into account becomes influential for the operation. Rabanser et al. (2019, 1–2, 5).

48 Sculley et al. (2014); Shendre (2020).

49 Bagchi et al. (2020).

50 Danzig (2018, 7); Greene (2006, 440–41).

51 For a discussion of the challenges of drift detection in a security context, see Nelson et al. (2014).

52 For example, in the early days of the COVID-19 pandemic, sudden systemic shifts in online consumer behaviour (such as drastic unanticipated spikes in searches for "face masks") stumped the predictive algorithms of several online retailers. Heaven (2020).

3. DEFINING KNOWN UNKNOWNS

Because of the inescapable causes of data issues discussed in Chapter 2, autonomous system failures are infinitely possible[53] – even, in a certain sense, inevitable.[54] Issues that cause failures come to be known[55] through the development of the weapon, the testing and legal review of that weapon, and that system's previous track record[56] (see figure 6). Modern militaries have a wide range of tools for quantifying and accounting for risks in complex systems,[57] and these sources of knowledge about the system will provide decision makers with ample information about the potential risks of deploying a given system in a given context.[58]

However, such issues are just a subset of all the actual issues a system might encounter in the real world.[59] It is impossible to know every single vulnerability that any given autonomous system might have or predict every single relevant data issue that such systems will encounter.[60] As such, all autonomous systems will be prone to inevitable

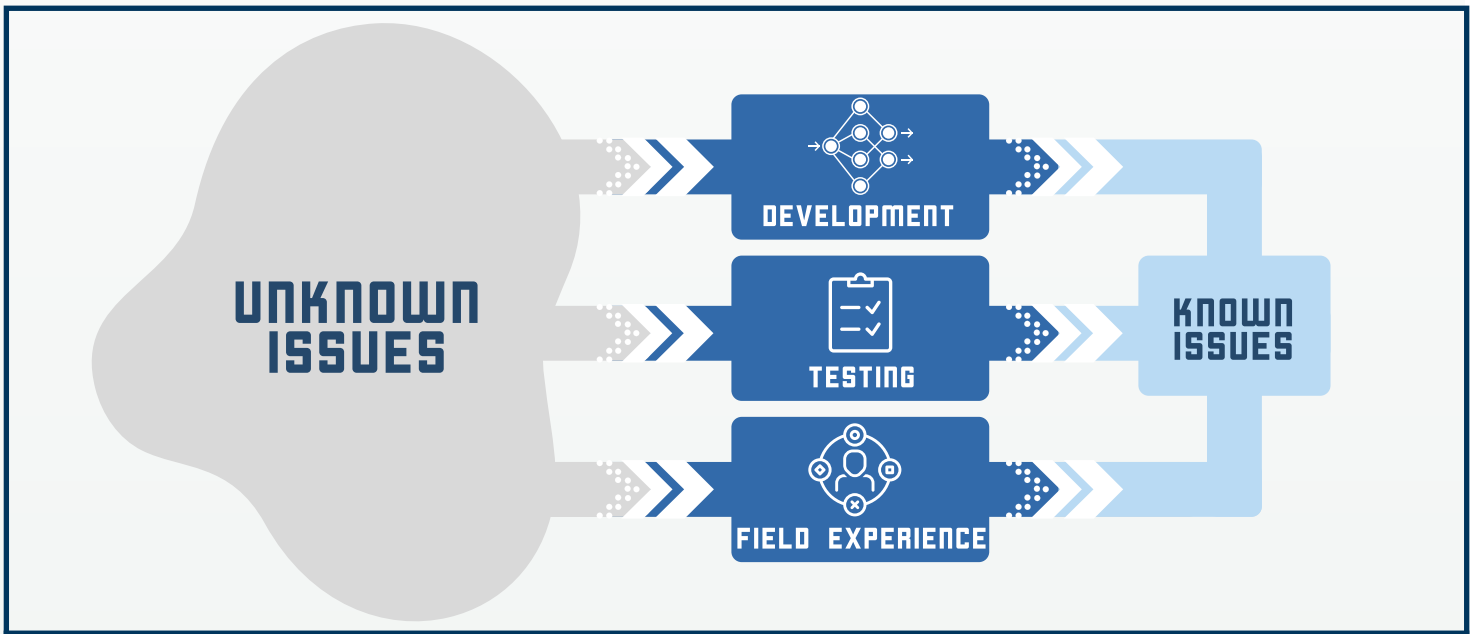


FIGURE 5. *Development, testing, and field experience make previously unknown issues known to the owner of autonomous systems. But these known issues are still only ever a subset of all the issues that are latent in that system.*

53 Carvin (2017, 2).

54 Or, put more bluntly, "An AI designed to do X will eventually fail to do X". Yampolskiy (2020). For a list of AI failures, many of which were sparked by issues described in Chapter 2, see Narayan Banerjee & Sekhar Chanda (2020); Yampolskiy & *Spellchecker* (2016, 1–2, 4–5).

55 For more on the idea of making "unknown" issues "known," see Kim (2012).

56 However, as Danzig (2018, 7) notes, this will not apply in cases of weapons that are rarely used.

57 See, for example, Department of the Army (2014). Non-military organizations also employ sophisticated risk assessment models for these kinds of purposes; see, for example, ECA (2019); MAA (2014).

58 Some militaries will also develop sophisticated tools to categorize these risks, along with extensive protocols to account for them.

59 All technical subject matter experts interviewed for this study affirmed that it is impossible to anticipate all potential AI failures by means of today's design, development and testing processes. A plurality of legal and policy experts similarly posited that existing review processes and mechanisms are likely not equipped, in their present configurations, to anticipate all errors.

60 Interview with Maria Vanina Martinez, 2 November 2020; interview with Davide Scaramuzza, 2 October 2020; interview with anonymous expert, 18 September 2020; interview with anonymous expert, 21 October 2020; interview with anonymous expert, 6 November 2020; Pinelis (2020).

accidents which cannot be foreseen.[61] Put simply, we know that such issues exist either now or will emerge in the future, but we cannot characterize or anticipate them.[62] One might call such data issues "known unknowns".

All complex weapon systems can have failure modes that cannot be foreseen. But it is likely to be harder to anticipate, quantify and characterize the risks associated with those issues in autonomous weapons. This is due to the inadequacy of present-day testing and verification processes for AI,[63] the difficulty of characterizing AI failure points, the low relative reliability of AI,[64] and the unpredictable conditions[65] and effects of autonomous system deployments.[66] As a result, it will be comparatively more challenging for militaries facing a complex conflict environment to determine whether and how data issues are likely to affect a deployed autonomous system and, by extension, where on the scale of reliability and risk that system will perform.[67]

61 See also Crootof (2016, 1373). For a similar formulation of this notion, see Morgan et al. (2020, 34). On the inevitability of accidents, see Scharre (2016, 5, 25); UNIDIR (2016).

62 Maas (2018, 2, 8) describes such accidents as "'unforeseeable' yet inevitable". See also Scharre (2016, 25); Yampolskiy (2020).

63 The traditional test, evaluation, validation and verification processes that grade the vulnerabilities, failure points and reliability metrics of complex non-autonomous systems are widely regarded as being inadequate for gauging the reliability, vulnerabilities and fit of complex autonomous systems. Interview with Tim McFarland, 13 November 2020; Haugh et al. (2018, 2-1 to 2-3); Luckcuck et al. (2019). Testing and risk assessment could be particularly challenging for those autonomous systems with a low level of understandability, those with a large number of tightly coupled interacting algorithmic components, or those with a low level of technical predictability. Flournoy et al. (2020, 7–10); NSCAI (2021, 137). Nor are these processes equipped to account for "data drift." Haugh et al. (2018, 2-3).

64 Lohn (2020a, 5). The development costs of raising the reliability of existing advanced AI models to match those of complex non-autonomous systems could be too high for many States; according to the Benaich & Hogarth (2020) the cost (in terms of computing time and power) of reducing a state-of-the-art image recognition system's failure rate by even few percentage points could run to many billions of dollars.

65 Non-autonomous systems are not subject to the "operational unpredictability" that makes it difficult to know in advance what inputs an autonomous system will encounter. A nuclear reactor is a complex, brittle system, but it is unlikely to encounter inputs for which it was not specifically designed; by contrast, an autonomous drone operating in an unfamiliar battlefield is quite likely to encounter inputs that cannot have been anticipated. Holland Michel (2020b, 5).

66 For example, the possible outcomes of a failure in an autonomous system – especially a system with highly autonomous capabilities and a wide "decision space" – could also be far more difficult to model than the outcomes of a malfunction in an equivalent non-autonomous system such as, say, a missile on a ballistic trajectory. Interview with Maria Vanina Martinez, 2 November 2020; Boulanin (2019, 20, 133); Boulanin & Verbruggen (2017, 70); Carvin (2017, 9); Defense Innovation Board (2019, 16, 66); Holland Michel (2020b, 5–7, 19); IEEE (2017, 128); Scharre (2016, 5); UNIDIR (2016).

67 Flournoy et al. (2020, 8); Jenihhin et al. (2019); Lohn (2020a, 5); Pinelis (2020).

4. LEGAL AND OPERATIONAL IMPLICATIONS

The vulnerability of all autonomous systems to data issues, and the relative unknowability of the failures those issues can lead to, has broad implications. The "comprehensive, context-based human judgment"[68] necessary for the legal use of autonomous weapons depends on the ability to anticipate and respond to data issues that could cause unintended harm. In this case, known unknown data issues raise the possibility of a legal paradox: Because such issues are unforeseeable with current testing and risk assessment measures, those employing these technologies may not be required to anticipate or respond to those issues. But because these issues are also by their very nature inevitable, these actors would need to address them, and could be held responsible for any harm stemming from a failure to do so. This chapter will explore these considerations in detail.

4.1 INTERNATIONAL LAW AND MILITARY AUTONOMOUS SYSTEMS

The body of international law governing hostilities does not make any specific provisions relating to digital data.[69] But the fact that data issues can lead autonomous systems to cause unintended harm[70] implicates a range of international legal principles.[71]

In practice, this means that if a data issue causes an autonomous weapon to inflict unintended harm, those that employed the weapon could bear responsibility for the harm if, among other factors, they:

› Could have reasonably anticipated the issue or did not account for the issue in their judgments related to proportionality and distinction;

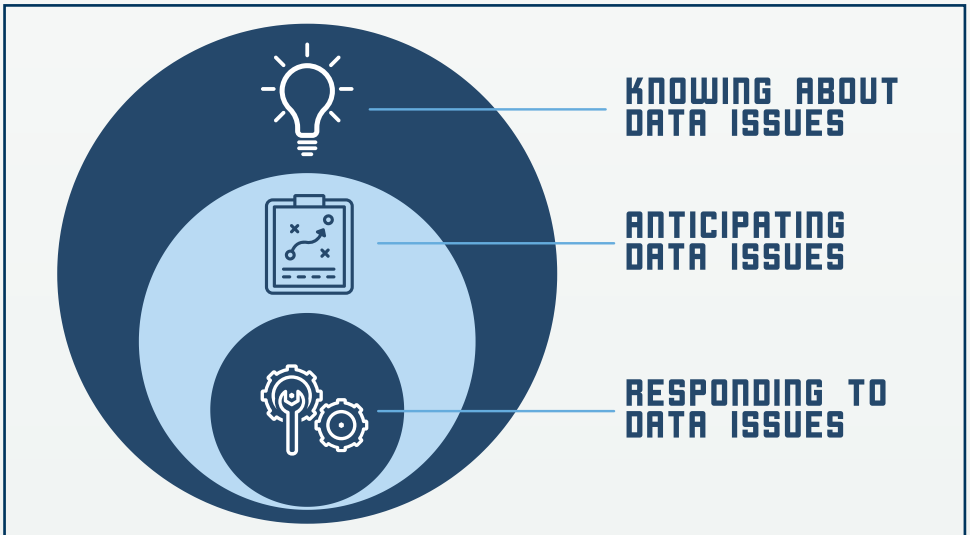


FIGURE 6. *In the application of international law to the use of autonomous weapons, users of such technologies would have to anticipate and (if possible) respond to data issues that might cause harm. Ultimately, doing so hinges on having prior knowledge of those issues.*

68 GGE on LAWS (2020, 5).

69 Interview with Rebecca Crootof, 18 November 2020; interview with Tim McFarland, 13 November 2020.

70 Including the misidentification of protected persons or objects, inadvertent escalation, friendly fire incidents, and other accidents. Scharre (2016, 5).

71 Namely distinction, proportionality, and precaution. Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Elisabeth Hoffberger-Pippan, 11 November 2020; interview with Molly Kovite, 11 November 2020; interview with Tim McFarland, 13 November 2020; interview with anonymous expert, 12 November 2020; Boulanin et al. (2020, 6).

- Did not maintain reasonable certainty until the end of the operation that the issue would not cause unintended harm; and
- Did not take all feasible measures to prevent that harm either during the lead-up to the attack or during the course of the operation.[72] (The exact threshold of "feasible" and the degree to which militaries must take these factors and risks into account will vary from case to case.[73])

This is also the case of unintended harm resulting from *adversarial* data issues.[74] However, if the adversarial action grants the adversary some level of control of the weapon, the legal responsibility to anticipate and prevent unintended harm would pass, either entirely or in part, to that adversarial actor.[75]

In other words, militaries would have to *anticipate, take into account, and where necessary respond* to all data issues in an operation—but only if they have reason to believe[76] that such issues could adversely affect their weapon systems and potentially cause unintended harm.[77] Therefore, the responsibility to address data issues in operations ultimately depends on whether the relevant parties *know* (or should have known) about those issues.

4.2 ANTICIPATING DATA ISSUES

If a particular type of data issue is known to cause failures in an autonomous system, those employing the system may have a responsibility to anticipate the likelihood of such issues in the environment[78] and factor those likelihoods into their decisions on whether and how to employ the weapon.[79] This responsibility to anticipate issues would likely be all the more crucial in cases where the user does not have the capacity to respond to issues in real time[80] (see Section 4.3). Many factors will impact the degree to which data issues can be anticipated.

- Some issues may be easier to anticipate.[81] Weather conditions, for example, can be measured and accurately forecast; intelligence collection can indicate the presence of enemy capabilities such as jamming weapons.[82] Other issues, such as the confluence of multiple distinct environmental factors, or previously unknown adversarial countermeasures,[83] will be more difficult to anticipate.

72 This was a shared view among all legal and policy subject matter experts interviewed for this study.

73 A broad range of factors contribute to the determination of whether a measure is feasible in any particular instance. Interview with anonymous expert, 12 November 2020; Quéguiner (2006, 809–11).

74 That is, if the possibility and effects of an adversarial attack on the autonomous weapon system could have been reasonably foreseen with the information available to the decision makers, and if the decision makers did not at the very least take that possibility into account when deciding to employ the system, some responsibility may apply. Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Elisabeth Hoffberger-Pippan, 11 November 2020; interview with anonymous expert, 12 November 2020.

75 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Molly Kovite, 11 November 2020.

76 Several States refer to this as the standard of "reasonably available" information; if information about, say, a system's potential for errors could not be reasonably discovered before the operation, it cannot necessarily be claimed that this information should have been known at the time of planning and executing the attack.

77 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Elisabeth Hoffberger-Pippan, 11 November 2020; interview with anonymous expert, 12 November 2020.

78 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Elisabeth Hoffberger-Pippen, 11 November 2020; interview with Molly Kovite, 11 November 2020; interview with Tim McFarland, 13 November 2020; interview with anonymous expert, 12 November 2020.

79 If there is a non-trivial probability of an issue arising in an operation, decision makers must weigh this factor against other factors that may be relevant to the decision, such as the density of civilian objects or persons in the area that may be at risk if the system fails, the military necessity of the objective, and the availability of other, potentially less risky means or methods of achieving the same goal. Per Quéguiner (2006, 796–800), belligerents are required to select the viable option that is least likely to cause injury to civilians.

80 Though such systems would be less vulnerable to adversarial data issues arising from communications jamming. Interview with Rebecca Crootof, 18 November 2020.

81 These may also be referred to as "planning assumptions". Boulanin et al. (2020, 9).

82 Interview with anonymous expert, 5 November 2020.

83 The fact that no autonomous system is "unattackable" also means that the possibility of adversarial countermeasures can never be entirely dismissed (except, perhaps, in instances where no active adversaries are operating), which may further undermine reasonable trust that a system will operate as intended in any given instance. Comiter (2019, 30). All weapons are subjected to countermeasures in war, but the effects of countermeasures on conventional weapons may be more predictable than the effects of countermeasures on AWS.

- Anticipating issues will be more challenging in missions with a long duration, operations over a wide area, operations in particularly complex or dynamic environments, and operations involving a highly active or sophisticated adversary.
- In cases of high "operational unpredictability",[84] it may be hard to anticipate not only *what* conditions a system might encounter but also *when* it might encounter them.[85]
- The amount of time and information that is available to make assessments prior to an attack will depend, in part, on whether the force is engaging in "deliberate" or "dynamic" targeting.[86]

If those using an autonomous weapon are unable to anticipate issues or conditions that could cause failures and harm, this inability would likely factor into the decision on whether a system could be launched.[87] It could be legally tenuous to employ an autonomous weapon where the chain of command has a limited sense of whether it will encounter known issues. (In many cases, this would also be operationally undesirable.[88]) These knowledge gaps may also confound decisions on how to apply measures to mitigate the harmful effects of those issues,[89] such as constraints to the system or tuning its parameters or functions.[90]

84 Holland Michel (2020b, 5).

85 Interview with anonymous expert, 6 November 2020.

86 Deliberate targeting generally requires a longer, more cautious and studied evaluation, whereas dynamic targeting may be carried out with less consideration of a wider set of facts. For a full breakdown of the steps that go into both deliberate and dynamic targeting, see Ekelhof & Persi Paoli (2020b). See also Devitt et al. (2020, 7).

87 Interview with anonymous expert, 15 October 2020; Boulanin et al. (2020, 7, 9); Quéguiner (2006, 796).

88 Interview with Molly Kovite, 11 November 2020; interview with Henrik Røboe Dam, 5 January 2020.

89 Bagchi et al. (2020).

90 Interview with Maria Vanina Martinez, 2 November 2020.

4.3 RESPONDING TO DATA ISSUES

In cases where data issues have a likelihood of manifesting, those employing autonomous weapons may have a responsibility to maintain some capacity to detect and respond to such issues to prevent harm.[91] The degree of this responsibility hinges, in part, on whether it is feasible to respond to such issues in sufficient time.[92] A range of factors will determine the degree to which operators can do so.[93]

- If an autonomous system channels a rich live feed (for example, video) to its human operators, they may have a higher capacity for detecting issues. In other instances, such as communications-denied environments, they may only have access to limited snippets of the weapon's incoming data.[94]
- If the live data collected by an autonomous system are complex and multifaceted, the operator's capacity to identify issues may be lower.[95]
- Certain issues may be imperceptible to the human operator. Spoofing might, for example, fool not only the autonomous system but also those who are overseeing it.[96]
- If a system has very low understandability,[97] it could be hard to find the cause of a problem

91 Interview with Anja Dahlmann, 11 November 2020; interview with Elisabeth Hoffberger-Pippen, 11 November 2020. This mirrors the practice of States that employ precision-guided weapons that can be controlled until moments before the impact. If operators of such weapons detect an issue, such as a civilian entering the blast zone of the weapon, they maintain the ability – and thus when feasible, the *responsibility* – to abort the weapon. Quéguiner (2006, 804). This is known as "shifting cold" or a "post-launch abort". Interview with Molly Kovite, 11 November 2020; Schmitt & King (2018).

92 Operators cannot be expected to abort a system if there is no reason to believe that a data issue is present. As one defence official commented in Reim (2020), "You can't just look at the outcome [from an AI system] and say, 'Well, the outcome wasn't what I expected, and therefore there must be something wrong with the system'".

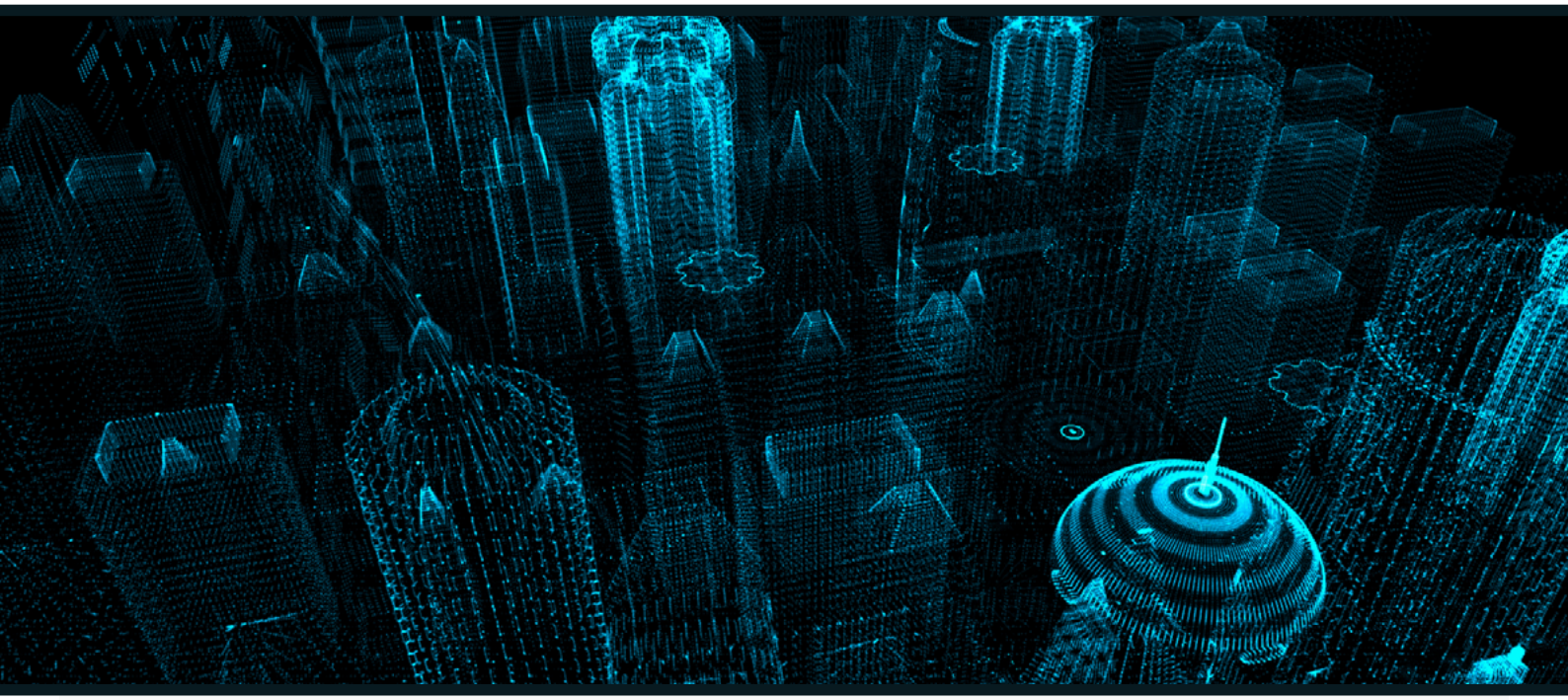
93 Interview with Anja Dahlmann, 11 November 2020; Interview with anonymous expert, 21 October 2020; interview with anonymous expert, 6 November 2020.

94 One anonymous expert, interviewed 6 November 2020, posited that given the difficulty of detecting issues in such arrangements, human operators may need to proactively check on a system periodically rather than simply wait to respond to issues when they arise.

95 Holland Michel (2020b, 17).

96 Interview with Henrik Røboe Dam, 5 January 2020; Bagchi et al. (2020); Lohn (2020b, 7–8, 11–12).

97 Interview with anonymous expert, 5 November 2020; Baksh (2020); Holland Michel (2020b, 15).



or know how to correct it,[98] even when it is evident that an issue exists.[99]

- Both insufficient trust and excessive trust ("automation bias"), could hamper operators' ability to recognize and correctly interpret certain issues.[100]
- In some cases, the form of human-machine interaction may not enable operators to intervene on a detected issue in enough time to prevent a failure.[101]

4.4 KNOWING ABOUT DATA ISSUES

Ultimately, the ability, *and thus responsibility*, to anticipate and respond to data issues that could cause unintended harm depends on whether those employing the system *know* about such issues. As described in Chapter 3, though States have a range of instruments to make weapon system vulnerabilities known prior to use, autonomous weapons can fail and cause unintended harm as a result of data issues that the decision makers employing those weapons could *not* have (and thus need not have) reasonably known about.[102] This could pose a legal paradox.

Unknown – If the review and risk assessment process for an autonomous weapon has given those employing the system no specific reason to believe that relevant data issues could emerge or result in unintended harm, they would not be required to take those specific data issues into account, or anticipate, detect or respond to those issues, during that mission.[103] In the event of harm arising from such "unknown" failures, these actors could not reasonably be accused of being responsible for failing to take the risk of such failures into account.[104] It would not be possible to prove that a human decision maker was criminally reckless or had wilful criminal intent in using a system in spite

98 European Commission (2020, 2).

99 Interview with Maria Vanina Martinez, 2 November 2020; interview with Davide Scaramuzza, 2 October 2020; Holland Michel (2020b, 17).

100 The issue of "trust calibration" is well studied in the literature on human-machine interaction. See, for example, CRS (2019, 31); Devitt (2018); Dietvorst et al. (2016); Hoffman (2017); Lewis et al. (2018); Morgan et al. (2020, 36); Parasuraman & Manzey (2010); Wang et al. (2016).

101 Boulanin et al. (2020, 19); Carvin (2017, 10).

102 In this context, an "accident" may therefore be characterized "as an undesired and unplanned *(but not necessarily unexpected)* event". Leveson (1995, 175), quoted in Scott & Yampolskiy (2019) [emphasis added].

103 This standard is sometimes framed as "reasonable foreseeability". Nohle & Robinson (2017).

104 Interview with Rebecca Crootof, 18 November 2020; interview with Molly Kovite, 11 November 2020. Describing unavoidable accidents in the context of non-autonomous weapons, Devitt (2021) refers to such failures as a "manifestation of an unforeseen uncertainty", which cannot be blamed on "human incompetence".

of that system's vulnerabilities, if those vulnerabilities were truly unforeseeable.[105]

Known – However, all autonomous systems will exhibit failures in use that the testing and review of those systems didn't identify. This fact could itself constitute a reasonable basis to doubt that unintended harm can be avoided.[106] If an autonomous system has known unknown failure modes, those deciding whether and how to acquire and use the system may not necessarily be able to claim to have "sufficient knowledge and understanding" of the system.[107] Nor could they necessarily claim to have been able to predict the exact effects of their decisions. The decision to acquire and employ an autonomous weapon in spite of this unquantified but nevertheless extant doubt[108] could leave parties open to responsibility for any resulting unintended harm.[109]

This discussion assumes that the exact source of errors can be identified after the fact,[110] whereas in reality the uninterpretability of complex autonomous systems could make it difficult to identify the exact source of errors – and if a weapon was fully autonomous and was destroyed in the attack, there may be no remaining evidence to indicate how or why it failed.[111]

4.5 EFFECTS OF THE KNOWN UNKNOWN PARADOX

This unusual facet of complex autonomous systems, which mirrors and potentially further confounds questions of how to assign human responsibility to unpredictable machine "decisions",[112] could lead to widely differing interpretations of how the law applies to these technologies.[113]

Those actors attempting to make a good faith assessment of the likelihood of unintended harm in the use of an autonomous system[114] will find little guidance on how to navigate the matter of known unknowns. For States that require their militaries to achieve a very high degree of certainty in operations, the difficulty of attaining certainty as to the possibility of data issues or their potential effects[115] would create de facto limits on where and how autonomous systems will be used[116] and on what type of weapons would be permitted to have autonomous capabilities.[117]

On the other hand, those with a looser threshold of required certainty in operations will have few compunctions about employing a potentially fallible system if that system's vulnerabilities are not specifically known. The unknowability of system errors may even provide convenient legal

105 Crootof (2016, 1375); Heyns (2014, 46).

106 Quéguiner (2006, 798) describes these types of situations as "case[s] of doubt" for which the belligerent would be required to obtain additional information prior to launching the attack. In 2019, the GGE on LAWS concluded that a lethal autonomous weapon "must not be used if it is…incapable of being used in accordance with the requirements and principles of [international humanitarian law]." GGE on LAWS (2019, 4).

107 The standard of "sufficient knowledge and understanding" is a widely shared view among parties to the GGE on LAWS. GGE on LAWS (2020, 8).

108 For example, GGE on LAWS (2020, 6) notes that a possible consensus recommendation from the GGE on LAWS is that "the use of weapon systems…that cannot *reliably or predictably* perform their functions in accordance with the intention of a human operator and commander to comply with [international humanitarian law] requirements and principles…is inherently unlawful" [emphasis added]. As Sassóli (2014, 324) notes, human decision makers do not need to understand all the technical details of their complex autonomous systems, but they must know the result of using that system.

109 Davison (2017, 17) describes how commanders could be held liable for deciding to employ an autonomous weapon if they cannot reasonably predict the effects of that employment, and if that employment then causes unintended harm.

110 The notion of "traceability" is key to many visions of responsible AI. Jobin et al. (2019).

111 Holland Michel (2020b, 15).

112 HRW & IHRC (2015); ICRC (2018, 14–16); Verdiesen et al. (2021).

113 Taken to its logical extreme, this paradox either implies that a) the use of any autonomous weapons that are *known* to have *unknown* data issues may always run afoul of international humanitarian law requirements or b) that there is no basis for responsibility for any unintended harm arising from any unanticipated issue – neither of which is practical. Interview with anonymous expert, 12 November 2020.

114 The idea of "good faith" is a key tenet of applying feasible measures (given the absence of hard lines relating to feasibility). See Lubell et al. (2019, 18); Quéguiner (2006, 810).

115 All of which contributes to the inherent unpredictability of autonomous systems. Holland Michel (2020b, 7).

116 Interview with Tim McFarland, 13 November 2020.

117 States might, for example, limit a system's weapons load so that the kinetic effect of any failure would be small.

cover against responsibility for harm that may not, in fact, have been totally accidental.[118] Even in cases where a system's specific failure points are known, detecting such issues in the environment might require detailed and extensive intelligence collection,[119] so much so that it could be claimed that such information is not reasonably available.[120] Similarly, though States can be held responsible for unintended harm arising from the use of weapons that have not been adequately tested,[121] it is hard to prove testing is "inadequate" if it fails to capture issues that could not have been specifically anticipated.

This could have troubling broader consequences. The adoption of unverified autonomous systems by risk-tolerant actors might, in turn, prompt more risk averse actors to employ their own unverified systems, since the operational need to defend against those systems could be deemed to outweigh the possibility of harms arising from known unknowns. This might lead to what some scholars have described as a dangerous "race to the bottom" on safety and security.[122]

118 This could contribute to what some groups have referred to as an "accountability gap" – a theory that decision makers and States may eschew responsibility for any "decisions" made by their autonomous systems. Chengeta (2020); Crootof (2016, 1366); HRW & IHRC (2015); ICRC (2014).

119 Interview with J.F.R. Boddens Hosang, 16 November 2020; Interview with anonymous expert, 5 November 2020; Interview with anonymous expert, 6 November 2020.

120 International humanitarian law does not require belligerents to possess highly sophisticated capabilities to collect such intelligence. Quéguiner (2006, 797).

121 Davison (2017, 16).

122 Scharre (2019, 15).

5. POTENTIAL SOLUTIONS

The factors described in the previous chapters indicate that, at current or near-future technology levels, extensive work is needed for militaries to achieve what one government report described as "justified confidence" when deploying an autonomous weapon.[123] The same will be true for those responding to autonomous weapons failures to assign proper responsibility for the harms they cause.

A range of technical approaches are often cited as potential solutions to prevent – or at the very least make known – the failures that data issues cause (see Annex I). However, while these approaches show promise, they are all still emerging research areas. And though they may reduce incidences of failure they may also increase the complexity of autonomous systems, thus creating new unknown vulnerabilities.[124] It is therefore safe to say that at least in the near- and mid-term future, technological solutions alone will not resolve the paradox of known unknown data issues.

This suggests that policy options may also be necessary to navigate the challenges discussed in this report. This chapter describes and discusses some of the most commonly referenced options. It finds that many of these, while potentially helpful, require significant additional research. Furthermore, none of these options is likely to fully address the problem of data issues if implemented in isolation.

5.1 FULL OR PARTIAL MORATORIUMS OR LIMITS ON USE

The ambiguity that could arise from potential data issues is part of the reason some States believe that lethal autonomous weapons must be prohibited or restricted.[125] However, such moratoriums face a challenge as there is still no widely agreed bright-line definition that would distinguish forbidden complex autonomous weapons from permissible automated systems.[126] Campaigns for a ban also face significant opposition from a variety of actors in the international debate.

Certain specific autonomous weapons capabilities, such as swarming systems that could display a high degree of unpredictability,[127] the application of autonomy to nuclear command and control, and systems with an "active learning" capability, are more widely agreed to be undesirable. But a consensus view on what would and would not fall under a potential moratorium or ban has yet to manifest itself either in the GGE on LAWS or within informal forums.

Others have proposed operational constraints to limit the possible effects of autonomous weapons accidents.[128] These include restrictions on anti-personnel systems[129] and constraints on where and when systems can operate to ensure that unanticipated failures would not cause harm.[130] Such constraints can be defined at the international level, or at the national level by way of rules of engagement or command decisions in battle.[131] However, constraints that are applied by technical

123 NSCAI (2021, 134).

124 Cummings (2020c, 6); Dahlmann & Dickow (2019, 12–13); Maas (2018, 3).

125 Interview with J.F.R. Boddens Hosang, 16 November 2020. Comiter (2019) suggests that in certain applications, vulnerability to attack may be a reason to never cede control entirely to the machine.

126 For example, the various existing automated weapons systems, such as close-in air defence weapons, that have already proven to operate with measurable reliability and relatively low rates of unintended harm. Interview with Henrik Røboe Dam, 5 January 2020.

127 IEEE (2017, 129).

128 GGE on LAWS (2020, 6, 9) describes this as a widely shared view among parties to the GGE on LAWS. See also, for example, Finland and the ICRC's submissions in GGE on LAWS (2020, 36, 88).

129 Arkin et al. (2019).

130 For example, barring autonomous systems from populated areas to ensure that if they do stray off course or malfunction catastrophically, the risk of collateral damage would be minimal.

131 At the national level, these constraints can be defined at various stages in the process leading up to the use of force. Ekelhof & Persi Paoli (2020b).

means could, themselves, be undermined by data issues.[132]

5.2 DIRECT HUMAN CONTROL

A plurality of technical, military and legal subject matter experts interviewed for this study described data issues as a foremost reason that humans should remain in the loop (at the tactical level) in critical operations involving autonomous systems. If States develop strict criteria for control of each kind of autonomous weapon under any given circumstance, operators could be positioned to get ahead of issues before harm arises, potentially even in cases where operators cannot identify the exact cause of the error. For example, operators might be trained to intervene if an autonomous system's "confidence score" for a particular task falls below a certain threshold.[133]

That being said, as discussed in Section 4.3, there are a wide range of factors that would undermine the extent to which human control could prevent critical failures. Furthermore, human control may fail to guarantee appropriate human responsibility for unintended harm involving autonomous systems in cases where data issues can be *claimed* as unanticipated.[134]

On the other hand, assigning total individual responsibility to human operators for *all harms* would likely lead, in certain instances, to misattribution of culpability.[135] In reality, the responsibility for harms that operators or their immediate superiors could not have prevented may lie further up the chain of command.[136] Distinguishing the categories of error that operators should be expected to prevent from those that they cannot (and that are therefore the responsibility of other actors, such as commanders, senior leadership or the manufacturer) remains an open research question.[137]

5.3 LIABILITY AND DUE DILIGENCE REGIMES

Regardless of who is deemed responsible for an autonomous weapon's harms, the international community has agreed that this responsibility must always be retained by humans.[138] Existing frameworks for assigning this kind of responsibility hinge on the notions of fault, recklessness, negligence (a failure to take due diligence) or, in the case of commander's responsibility, on whether the person had "reason to know" that their subordinate would behave in a certain way. But it is difficult to prove any of the above in cases in which the harm in question can be claimed to have been unforeseeable.[139]

Some commentators have therefore proposed a "strict liability" framework that would render States fully liable for any unintended harm resulting from operations involving autonomous weapons, regardless of the finding of fault.[140] Such an arrangement would compel States to take both *known* data issues *and* the possibility of *known unknown* data issues into account in the development, review, and use of autonomous weapons. Strict liability regimes are already applied to certain dangerous and unpredictable activities,[141] as well

132 For example, a system programmed to abide geographical constraints would rely on navigational GNSS (Global Navigation Satellite System) data (see Annex II) to estimate its location; any interference in that navigational data feed would undermine the system's ability to adhere to its constraints. An anti-material-only constraint will only be effective if the system in question has the ability to detect and avoid personnel in its data feeds.

133 Interview with Anja Dahlmann, 11 November 2020.

134 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with anonymous expert, 21 November 2020.

135 Interview with Rebecca Crootof, 18 November 2020; Elish (2019, 55).

136 For example, responsibility may lie with the commander who opted to deploy an autonomous system without a solid grasp of the likelihood of failure. Davison (2017, 17). Human decisions factor into all stages of the process leading up to the use of force; see Ekelhof & Persi Paoli (2020b).

137 GGE on LAWS (2019, annex IV) has stressed the principle that human-machine interaction (HMI) must ensure compliance with international law, but the technical particularities of designing effective HMI, as well as the attendant policies to accompany HMI arrangements, have yet to be resolved.

138 GGE on LAWS (2019, annex IV).

139 For an overview of the potential challenge of applying existing criminal liability regimes to autonomous weapons systems, see Crootof (2016, 1375–81); Mann (2019).

140 This proposal was argued first and most extensively by Crootof (2016). European Commission (2020, 12–16) discusses the various factors implicated in liability matters related to AI-enabled systems and proposes strict liability as a potential legal measure. See also Geiß (2017).

141 Such as owning wild animals or demolishing buildings. Interview with Rebecca Crootof, 18 November 2020; Crootof (2016, 1395).

as international treaties relating to accidents in outer space,[142] and are a common proposed legal option for self-driving vehicles.[143]

Absent a strict liability regime, establishing clearer lines of responsibility for autonomous weapons failures would likely, at a minimum, rely on a tailored set of due diligence criteria. These criteria would probably have to extend upstream from the system's end user to include the manufacturer of the system,[144] and possibly also downstream to include extensive intelligence collection of the battlespace. Given the complexity and changeability of data issues, such regimes of supervision would need to be continuous, comprehensive, dynamic, and meticulously tracked, and be applied at every stage of the life cycle, starting long before deployment and continuing through assessment and evaluation after all instances of use.[145] More work is needed to develop and standardize such regimes.

Finally, achieving greater clarity with regard to responsibility and liability will probably require a closer, science-based examination of the root causes and effects of data issues, and perhaps a corresponding classification schema for AI accidents.[146] This kind of schema could help parties distinguish known unknown accidents from *unknown* unknown accidents, and apportion responsibility accordingly.

One way or another, given that autonomous systems accidents arise from the interaction of the technology with its environment, its users and the regulations and norms under which it is operated, such liability or due diligence frameworks would likely have to account for legal and societal dimensions in addition to technical aspects.[147]

5.4 LEGAL REVIEWS

An often-cited measure to help decision makers address the ambiguity arising from data issues is the legal review process.[148] This process could help ensure that as few as possible of the issues that emerge during an operation are *unknown*. To achieve this effect, a review would have to:[149]

- Identify any data issues (including adversarial data threats[150] or the risk of "emergent effects"[151]) that would undermine the application of relevant laws in any attack.[152]
- Determine whether the system's training and testing environments closely match the proposed operational environments.[153]
- Measure and validate the reliability of the system.[154]
- Evaluate the degree to which the human element in the planned uses of the system could reliably anticipate or respond to issues that arise.[155]

Based on these findings, reviewers could create guidelines to ensure that the system never encounters these issues or is never used in environments or in ways that would present it with inputs that are significantly different from those inputs

142 Geiß (2017).

143 Evas (2018).

144 Cummings (2019).

145 Margulies (2019, 19–22).

146 See, for example, Scott & Yampolskiy (2019).

147 Verdiesen et al. (2021).

148 A plurality of legal experts consulted for this study cited legal reviews as being elemental for forestalling failures arising from data issues. Parties to the GGE on LAWS also widely share this view; see GGE on LAWS (2020, 5).

149 For a detailed overview of the likely elements of legal reviews for an autonomous weapon, see Lewis (2019).

150 Farrant & Ford (2017, 411–12).

151 Defense Innovation Board (2019, 13); Ilachinski (2017). See also European Commission (2020, 9).

152 For example, issues that would result in the weapon causing indiscriminate harm. Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Rebecca Crootof, 18 November 2020; Boulanin (2015, 14–15); Margulies (2019, 5).

153 Interview with Molly Kovite, 11 November 2020; interview with Tim McFarland, 13 November 2020.

154 Farrant & Ford (2017, 410) and ICRC (2006, 946).

155 Interview with J.F.R. Boddens Hosang, 16 November 2020; Haugh et al. (2018, 3-2).

it was developed for.[156] At a minimum, reviewers could establish guidelines to ensure that if the weapon is used in such environments or ways, the possibility of failure is taken into account in the decision to launch or control it.[157]

Considerable further work is necessary to make such reviews possible. Above all, existing testing and evaluation techniques for complex systems will need to be significantly overhauled to identify vulnerabilities, rate their likelihood, quantify a system's reliability, and ensure that the mode of human-machine interaction would enable operators to respond to issues appropriately.[158] Given the technical complexity of the material, these reviews are also likely to require closer input from engineers than is ordinarily needed for non-autonomous weapons.[159] In the absence of such measures, legal reviews may only be able to certify systems for use in extremely narrow, closely supervised circumstances, or circumstances where the likelihood of unintended harm arising from any kind of failure would be exceedingly low.

5.5 RECURSIVE TESTING AND REVIEW

Because unanticipated autonomous system failures are inevitable, numerous parties to the debate have called for a recursive testing and review process under which any previously unknown data issues trigger a new review or testing process.[160] This would enable States to implement technical fixes, refine parameters and guidelines of use, or at least ensure that decision makers take the potential for such failures into account in any future operations.[161]

But it may sometimes be impractical to retrain or recode and re-verify a system every time an issue is discovered, especially if such issues emerge with high frequency.[162] Furthermore, any technical modification could, in turn, introduce new vulnerabilities that could themselves cause failures.[163] The same could also be true for changes to how the system is used. As a result, these modifications may also have to be subject to testing, review or risk assessment.

156 For instance, if a review ascertains that a particular type of jamming may cause a weapons system to misidentify targets, evaluators may draft internal rules requiring decision makers to abstain from the employment of that weapon in any instance where such jamming may be present. Or if a system is exclusively trained and tested against uncluttered environments with high visibility, reviewers might prohibit the use of that system in cluttered environments with degraded visibility. European Commission (2020, 8) suggests similar measures for evaluation of civilian AI.

157 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Molly Kovite, 11 November 2020.

158 Christie (2020); Defense Innovation Marketplace (n.d.); Flournoy et al. (2020); Boulanin & Verbruggen (2017, 70); Defense Innovation Board (2019, 16, 66); Herpig (2019, 35); Koopman & Wagner (2016). One relevant research thread seeks to better define the factors that determine how well-matched a proposed operational environment is to the system's design environment, as well as the factors that may affect the variability of environments and the degree to which they may manifest relevant data issues. Koopman & Fratrik (2019) propose one such framework for autonomous vehicles. Ad Hoc ALFUS Working Group (2007, 30-35) proposes a range of potentially relevant elements of a framework to categorize environmental complexity for unmanned vehicles. See also Jenihhin et al. (2019, 4).

159 Interview with J.F.R. Boddens Hosang, 16 November 2020; IEEE (2017, 118).

160 European Commission (2020, 7–8) proposes recursive reviews as a potential measure to address AI-enabled systems that experience changes unforeseen by their manufacturer. Schmidt (2021).

161 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Rebecca Crootof, 18 November 2020; interview with Molly Kovite, 11 November 2020.

162 Cummings (2020a, 3) describes this approach as a "finger in the dyke" solution. Maria Vanina Martinez, interviewed 2 November 2020, cautioned that merely identifying the underlying cause of errors may be a challenge, especially if a system is uninterpretable.

163 Interview with Maria Vanina Martinez, 2 November 2020; interview with Davide Scaramuzza, 2 October 2020. Sculley et al. (2014, 2) refer to this as the "CACE Principle: changing anything changes everything."

5.6 STANDARDS AND KNOWLEDGE-SHARING

A plurality of experts interviewed for this study advocated for knowledge-sharing and common technical standards to help States navigate the challenges of implementing the above solutions.

International standards are a commonly proposed tool for enabling the responsible use of autonomous weapons.[164] Such standards could extend beyond the literal technical elements of autonomous systems, to include the arrangements of human interaction and control that will play a vital role in enabling States to anticipate and respond to data issues.[165] However, standards development for autonomous systems remains relatively nascent. Many of the most advanced initiatives are intended for civilian autonomous systems, which do not pose all the same challenges as military weapons.

To aid in the validation process for machine-learning-based systems, in particular, some have also proposed sharing standardized data sets that would guarantee a common baseline of robustness for all systems operating in a specific role or environment. Some experts have noted that States or system vendors may be reticent to share data sets, given the sensitivity of their content (much military data is secret) and the advantage that more robust AI might provide to adversaries.[166]

A similar option might include the independent certification of data sets to ensure they sufficiently capture a system's proposed operating environment. Such validation processes could consider not only the size and source of data sets but also the degree to which they are representative of the target environment (to avoid bias) and integrity (so as to ensure that the data have not been poisoned). Open criteria for data sets have not yet been established for military-type data, and some types of data set may be harder to certify than others.[167]

The sharing of non-technical know-how and resources could also be valuable. For example, a number of militaries are already building autonomous systems doctrine to reduce the uncertainties posed by data issues, leveraging their extensive experience in risk management for other complex systems such as aviation. Enabling access to these resources and expertise could help broaden the adoption of rigorous risk-reduction policies and strategies among a broader and more diverse cross-section of states.

164 For a discussion of technical standards for autonomous systems and their potential applicability to autonomous weapons, see ICRC (2019, 21–24).

165 Daiki (2020).

166 Interview with Elisabeth Hoffberger-Pippan, 11 November 2020; interview with Tim McFarland, 13 November 2020; Mulchandani (2020); Pinelis (2020).

167 Interview with Rebecca Crootof, 18 November 2020.

CONCLUSION: FIVE AVENUES FOR ACTION

The issues described in this report are formidable and, in some cases, inherent. But the risks of inaction are potentially grave. The following five avenues for action could bolster efforts to minimize the risks of unintended or unaccountable harms arising from the use of military autonomous systems. Like all international initiatives relating to autonomous military systems, they will require close cooperation between stakeholders from all domains, including governments, militaries, civil society, academia and the technology sector.

1. **Perform advanced, collaborative research on the legal review process.** Legal reviews are likely to be key to addressing data issues. Developing legal review procedures that resolve the many ambiguities described in this report will require significant new research, collaborative dialogue and knowledge-sharing.

2. **Develop classification criteria for data issues and resulting failures; specifically, develop criteria to distinguish *known unknown* issues from *unknown unknown* issues, and frameworks to assign appropriate responsibility in cases of harm arising from such issues.**[168] A finer-grain scheme for differentiating between different types of failure[169] – and a clearer framework designating the actors for whom those failures should be knowable – could aid efforts to quantify risk in operations and assign due responsibility for unintended harm arising from data issues.

3. **Share specific knowledge on technical and normative approaches to data and risk in relation to autonomous military systems.** Given the formidable challenge of characterizing data issues, to say nothing of addressing them through technical approaches, all stakeholders should be encouraged to share knowledge across political and disciplinary divides. This especially applies to sharing of best practices, given that even good faith efforts to minimize the risks of data issues in autonomous systems could be frustrated by the complexity and ambiguity of data issues.[170] A number of militaries already possess significant shareable relevant knowledge (for example, sophisticated risk assessment tools and procedures) that could serve as a foundation for assessing autonomous systems risks; the distribution of these resources would be beneficial for all actors seeking to mitigate the risks of autonomous systems.

4. **Study adversarial measures and their effects on autonomous weapons.** No autonomous system is "unattackable",[171] and many of the most dangerous and unpredictable data issues for autonomous systems could arise from adversarial actions. By foregrounding the science of adversarial measures, the international community will better place itself to model their effects and, as necessary, take adversariality into account in the development of norms or policies for the development and use of autonomous systems.

5. **Adopt a system-of-systems approach to studying data issues.** Failures in autonomous systems arise from the interaction of a range of subsystems: not just sensors and algorithms but also actuators, power sources, communications devices and other systems in the battlespace. Taking all these interacting systems into account will help guide parties to more grounded solutions than discussions that solely focus on the algorithmic element of autonomous technologies.

168 With gratitude to Rebecca Crootof for input on this recommendation.

169 Scott & Yampolskiy (2019).

170 Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Rebecca Crootof, 18 November 2020; interview with Anja Dahlmann, 11 November 2020; interview with Molly Kovite, 11 November 2020; Boulanin (2015); GGE on LAWS (2019, 3).

171 Comiter (2019, 30).

ANNEX I: TECHNICAL APPROACHES TO ADDRESS DATA ISSUES

A wide range of technical approaches are under development to address the types of data issue described in this report. This Annex describes these approaches and discusses their respective readiness level and the challenges that may be associated with their implementation.

AUTOMATED DATA ISSUE DETECTION

Many accidents arising from data issues could potentially be avoided by endowing autonomous systems with the capability to detect when they are encountering an input for which they were not designed or validated. A growing body of research seeks to develop "out-of-distribution detection," "anomaly detection," and "shift detection" features that can endow AI systems with this capacity.[172] A related strand of emerging research seeks to develop tools for quantifying the degree to which the given environment matches the environments for which the system was developed, trained and validated.[173] Such features might enable systems to revert to a fail-safe mode when a data issue is encountered or enable decision makers to develop a more informed understanding of the likelihood or risk of failures before launching a system.[174]

Confidence scores, which are employed in some machine-learning systems to indicate the degree to which the characteristics of the given data point resemble the characteristics of the training inputs that were marked with the same label, may provide some indication that a system is encountering an unusual input. As such, these features could serve as a de facto form of anomaly detection. An autonomous system could be coded, for example, to revert to a fail-safe mode if its confidence score falls below a particular threshold. But relying on confidence scores is not necessarily, at current technology levels, a foolproof option. A growing body of research has demonstrated that systems can be liable to produce high-confidence erroneous outputs when they are fed certain anomalous or adversarial inputs.[175]

HYBRID SYSTEMS

Sometimes, data issues cause failures in autonomous systems because they lack the capacity to reason through those issues and determine an appropriate course of action. One approach that seeks to correct this problem is hybrid intelligent systems that integrate machine learning with symbolic (also known as rule-based or knowledge-based) algorithms. Such hybrid systems, which are also known as neurosymbolic systems, pair machine learning's capacity to accurately classify incoming data with symbolic AI's capacity to draw logical conclusions from those outputs. In other words, such systems combine the "cognitive" capacity of machine-learning systems with the "logic-based" capacity of symbolic AI.[176]

For instance, a neurosymbolic system designed for tracking objects through physical space would employ a learning-based vision algorithm to identify and follow the object, and a symbolic algorithm to validate the object's behaviour against a library or model of the expected physical behaviours of such objects. By cross-checking the output from the learning system against the library or model, this secondary system could deduce that when the object disappears from view it is because it has passed behind another object, rather than because it no longer exists. Hybrid systems could also theoretically detect when an object is exhibiting unexpected or unnatural behaviour.[177]

172 Interview with anonymous expert, 15 October 2020; Amodei et al. (2016); Lohn (2020a, 5–6); Rabanser et al. (2019, 1–2, 5); Ren & Lakshminarayanan (2019).

173 Strout (2020).

174 For example, see challenges discussed in Bulusu et al. (2020, 132, 343–44) with respect to out-of-distribution detection for deep learning.

175 For example, Nguyen et al. (2015, 7–8).

176 Interview with Maria Vanina Martinez, 2 November 2020.

177 Smith et al. (2019, 3–6).

Such an architecture could prevent autonomous systems from making an erroneous "best guess" output based on an anomalous or incomplete data input.[178] However, hybrid AI, like anomaly detection, remains an emerging research space that has for the most part only been demonstrated in experimental settings. Additionally, to be effective in highly complex and dynamic environments, hybrid systems may require extremely detailed and complex logical models that are challenging to build and require detailed domain-specific information about the proposed operating environment.[179] These data may not always be available in a conflict setting.

MULTISENSOR SYSTEMS

To account for the natural limitations of all data types and the sensors that collect them, one common approach is to employ multiple data sources aboard a single system to cross-validate the attributes of any given observed object or phenomenon in the environment.[180] For example, an autonomous system might employ both a vision system *and* a radar system to confirm the existence and identity of physical objects in the area of operations. If the vision-based algorithm drawing data from the camera incorrectly identifies a cloud as an enemy aircraft, or if the image recognition system is being spoofed by an adversarial example, the radar could demonstrate that no aircraft is, in fact, present.[181]

Sensor fusion can also generate a more granular identification of the observed object or phenomenon, thus reducing the effect of data quality issues that could arise in a standalone sensor. For example, it might be challenging for a target recognition system to deem whether a vehicle is part of an adversary force based on infrared data alone, but if a secondary hyperspectral sensor detects that the system is emitting chemical signals associated with explosive ordinance, it may be able to achieve a justifiably higher confidence identification.[182]

Fusion is a common technique for automated physical systems. Self-driving vehicles rely on sensor fusion to build a comprehensive and reliable model of their surroundings.[183] However, as the ongoing challenges of achieving reliable, fully automated self-driving vehicles demonstrates, sensor fusion alone cannot prevent all data issues, nor does it necessarily mitigate issues arising from discrepancies between the operational environment and the system's training and development.

Furthermore, the addition of new sensing modalities to an autonomous system increases the challenge of comprehensive testing and validation.[184] Additionally, the limited power and payload capacity of autonomous systems naturally constrains the number and variety of sensors they can carry.

EXPLAINABLE AI

Data issues can be especially difficult for human decision makers to anticipate or detect in uninterpretable "black box" autonomous systems. Uninterpretable AI also poses a challenge for identifying issues in development and testing.[185] A broad body of research seeks to develop AI systems that are either inherently transparent or that are complemented by tools that "explain" how the system works. While much hope has been placed in explainable AI research, this remains an emerging field that has not as yet achieved provable, rep-

178 Meyer-Vitali et al. (2019, 10–18).

179 Interview with Maria Vanina Martinez, 2 November 2020.

180 Interview with Davide Scaramuzza, 18 September 2020; interview with anonymous expert, 15 October 2020; interview with anonymous expert, 6 November 2020.

181 For example, an inertial navigation system – which does not rely on GPS – could serve to cross-validate a system's GPS readings; the system could be coded to revert to a fail-safe mode whenever the two readings diverge, since this may be a sign that the GPS is being spoofed. Interview with Henrik Røboe Dam, 5 January 2020.

182 Intelligence "fusion" is a common practice in modern (non-automated) targeting operations. In many cases, operators are required to validate one source of data against a second source. Interview with J.F.R. Boddens Hosang, 16 November 2020; interview with Molly Kovite, 11 November 2020; interview with Henrik Røboe Dam, 5 January 2020; interview with anonymous expert, 5 November 2020.

183 Cohen (2018).

184 Cummings (2020c, 3, 6).

185 Boulanin & Verbruggen (2017, 70); Hagström (2019, 37); Haugh et al. (2018, 3-2); Ilachinski (2017, 199–209).

licable solutions for the complex challenges described in this report.[186]

Larger training and testing data sets

At the moment, the only consistently and demonstrably reliable method to ensure that machine learning systems are validated against the widest possible degree of variance in data[187] is to increase the size of the data sets on which they are trained and tested.

However, creating data sets that comprehensively cover the variability of a complex real-world uncontrolled environment is a resource-intensive process[188] involving a lengthy list of steps. These include cleaning, curation, optimization and labelling processes that require significant time, personnel and expertise.[189] As such, the creation of sufficiently comprehensive data sets, at least today, would likely be prohibitively difficult and expensive for most militaries.[190]

Though certain types of data for training machine-learning systems are widely and easily accessible – such as image recognition data – these generally only encompass civilian environments and applications, meaning that they would not make military autonomous systems robust against conditions specific to warfare.[191] And even large, tailored data sets can only systematically capture those data issues that developers and evaluators of an autonomous system know in advance to be potentially problematic. ("One shot" learning, a technique for training AI systems on a very limited data set, has shown some early promise, though it likely would not be viable for highly variable environments such as conflict settings, at least for the foreseeable future).

Synthetic training and testing

Because the creation of large custom data sets is difficult and expensive, an increasingly popular option for widening the development environment of AI is the use of synthetic data.[192] For example, a machine-learning vision system for identifying airplane types in satellite imagery might struggle to identify aircraft that were only pictured a handful of times in the testing data; by training and testing such a system with synthetic images of those aircraft types, it might achieve a more robust performance.[193] Synthetic data can also randomize various qualities of the input data, thus increasing the diversity of inputs that the system is trained against and reducing the likelihood of edge cases.[194]

Another potential technique for expanding the training envelope of machine-learning systems is the use of generative adversarial networks, an architecture that generates a large number of input examples derived from an original set of real training data. By generating a wide variety of examples, generative adversarial networks can be used either to generate highly realistic synthetic training data (which can subsequently be used to train a model) or to directly train the model itself.[195]

However, because they are artificially generated, such synthetic worlds do not yet fully or perfectly replicate the complex chaotic physics of real-world environments[196] and may fail to properly capture the shifts or one-off edge cases that emerge in real-world data.[197] The uninterpretability of synthetic data creation engines would also

186 Holland Michel (2020b, 22); Rudin (2019); Schmelzer (2019).

187 Taori et al. (2020, 2).

188 Flournoy et al. (2020, 9)

189 For a description of the "data pipeline" process for machine-learning systems, see Chahal et al. (2020, 5); Stumborg (2019).

190 Interview with Rebecca Crootof, 18 November 2020; Cummings (2020a); Svenmarck et al. (2018, S-1 to 5-4).

191 Chahal et al. (2020, 10–11); Freedberg (2019).

192 Beery et al. (2020); Shakhuro et al. (2019).

193 Freedberg (2020); Shermeyer (2019).

194 See for example, Tremblay et al. (2018).

195 Interview with Maria Vanina Martinez, 2 November 2020; interview with anonymous expert, 21 October 2020; Creswell et al. (2018, 62–64). Generative adversarial networks could also be used to generate difficult-to-detect adversarial examples for spoofing AI systems. Tucker (2019).

196 Chahal et al. (2020, 11). Such simulations will be "limited by the scope of the model", per Hagström (2019, 38).

197 Interview with Davide Scaramuzza, 18 September 2020; Taori et al. (2020).

pose a challenge to the validation of any systems trained on those data. As a result, synthetic data alone cannot be expected to resolve the inherent challenges that may face deployed systems that have had a limited training environment.[198]

TRANSFER LEARNING AND DOMAIN ADAPTATION

A more fundamental field of research seeks to address the source of the problem of discrepant data by developing machine-learning systems that are robust in environments that differ from the environments for which they were trained and tested. Such techniques, which are known broadly as "domain adaptation" and "transfer learning",[199] could significantly improve the robustness of learning-based systems, but they remain an emerging area of research and have not yet been thoroughly demonstrated for complex, dynamic, uncontrolled environments[200] or military settings.

198 Dahlmann & Dickow (2019, 13).

199 For an introduction to domain adaptation and transfer learning, see Lohn (2020a, 6).

200 Kouw & Loog (2018, 29).

ANNEX II: SENSORS IN FOCUS

Physical autonomous systems will collect data through a variety of sensing systems. The most common sensing modalities that autonomous weapons will employ include:

- **Electro-optical cameras,** which collect imagery in the visible light spectrum. Stereoscopic cameras with multiple lenses enable the estimation of depth in physical space.
- **Infrared cameras,** which detect objects and phenomena based on variations in temperature
- **Hyperspectral sensors,** which generate imagery by detecting objects or phenomena across a wide band of the electromagnetic spectrum
- **GNSS (Global Navigation Satellite System) receivers, such as GPS and GLONASS,** which take in information about the system's geographic location from satellites
- **Inertial measurement units,** which collect information about a system's orientation, force and movement through space
- **Radar,** which detects objects by emitting radio pulses into the environment and recording returned pulses that bounce off those objects
- **Lidar,** which detects objects by emitting laser pulses into the environment and recording returned pulses that bounce off those objects
- **Acoustic sensors,** which collect sounds from the environment
- **Radio receivers,** which collect communications data such as remote-control links or intercepted "signals intelligence" (for example, intercepted radio chatter or tracked cell phone location data)
- **Sonar,** which is used to detect objects under water by either recording their emitted sounds or receiving returned sound pulses that bounce off the object

Just like the five human senses, all sensors used by autonomous systems today have limitations that give rise to incomplete, poor quality or false data inputs. For example, electro-optical cameras cannot operate in low-light conditions, can sometimes fail to detect surfaces that have a uniform texture (such as a clean painted wall)[201] or differentiate between surfaces that have similar colour or lighting,[202] and may be thrown off by glare or reflections. Acoustic sensors have a limited range, particularly in noisy environments.[203] Lidar tends to struggle in adverse weather.[204] Radar systems designed for traditional air defence may not be able to detect small, distant, slow-moving objects.[205]

All such systems could also be susceptible to adversarial action. While many of the most famous examples of adversarial AI attacks focus on vision-based object recognition systems, a growing body of research has pointed to the susceptibility of a broad range of sensor and receiver types commonly used in autonomous systems, including lidar and radar.[206]

201 Interview with Davide Scaramuzza, 2 October 2020; interview with anonymous expert, 6 November 2020.

202 For example, this effect contributed to the fatal crash, in 2016, of a Tesla Model S that was unable to distinguish the side of a truck from the sky above.

203 Holland Michel (2020a).

204 Interview with Davide Scaramuzza, 2 October 2020; Yurtsever et al. (2020, 11).

205 Alhaji Musa et al. (2019).

206 See Cao et al. (2019); Yeh et al. (2020).

BIBLIOGRAPHY

Ackerman, Evan. 2016. "Fatal Tesla Self-Driving Car Crash Reminds Us That Robots Aren't Perfect." IEEE Spectrum, 1 July, 8.00 p.m. GMT. As of 16 January 2021: https://spectrum.ieee.org/cars-that-think/transportation/self-driving/fatal-tesla-autopilot-crash-reminds-us-that-robots-arent-perfect

Ad Hoc ALFUS Working Group. 2007. *Autonomy Levels for Unmanned Systems (ALFUS) Framework*. Washington, DC: National Institute for Standards and Technology.

Aisch, Gregor, Josh Keller & Sergio Peçanha. 2016. "How a Cascade of Errors Led to the U.S. Airstrike on an Afghan Hospital." *New York Times*, 29 April.

Alhaji Musa, Surajo, Raja Syamsul Azmir Raja Abdullah, Aduwati Sali, Alyani Ismail & Nur Emileen Abdul Rashid. 2019. "Low-Slow-Small (LSS) Target Detection Based on Micro Doppler Analysis in Forward Scattering Radar Geometry." *Sensors (Basel)* 19 (15): 3332. doi:10.3390/s19153332

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané. 2016. "Concrete Problems in AI Safety." arXiv:1606.06565v2.

Arkin, Ronald, Leslie Kaelbling, Stuart Russell, Dorsa Sadigh, Paul Scharre, Bart Selman & Toby Walsh. 2019. "A Path Towards Reasonable Autonomous Weapons Regulation." IEEE Spectrum, 21 October, 3.50 p.m. GMT. As of 9 December 2020: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/a-path-towards-reasonable-autonomous-weapons-regulation

Artner, Stephen, Richard S. Girven & James B. Bruce. 2016. *Assessing the Value of Structured Analytic Techniques in the U.S. Intelligence Community*. Santa Monica, CA: RAND Corporation.

Athalye, Anish, Logan Engstrom, Andrew Ilyas & Kevin Kwok. 2017. "Fooling Neural Networks in the Physical World with 3D Adversarial Objects." Labsix, 31 October. As of 15 January 2021: https://www.labsix.org/physical-objects-that-fool-neural-nets

Atyabi, Adham, Somaiyeh MahmoudZadeh & Samia Nefti-Meziani. 2020. "Current Advancements on Autonomous Mission Planning and Management Systems: An AUV and UAV Perspective." *Journal of Annual Reviews in Control* 46: 196–215. doi:10.1016/j.arcontrol.2018.07.002

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. "Vision Paper: Grand Challenges in Resilience: Autonomous System Resilience through Design and Runtime Measures." *IEEE Open Journal of the Computer Society* 1: 155–72. 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 December 2020: https://www.nextgov.com/cybersecurity/2020/07/artificial-intelligence-systems-will-need-have-certification-cisa-official-says/166600

Barcelona Declaration. n.d. "Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe." As of 21 February 2021: https://www.iiia.csic.es/barcelonadeclaration

BBC. 2020. "Uber's Self-Driving Operator Charged over Fatal Crash." 16 September. As of 27 November 2020: https://www.bbc.com/news/technology-54175359

Beery, Sara, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Markus Meister, Neel Joshi & Pietro Perona. 2020. "Synthetic Examples Improve Generalization for Rare Classes," paper, *IEEE Winter Conference on Applications of Computer Vision*, Aspen, Colo., 1–5 March 2020.

Benaich, Nathan, & Ian Hogarth. 2020. *State of AI Report*. As of 13 April 2021: https://www.stateof.ai.

Bhambri, Siddhant, Sumanyu Muku, Avinash Tulasi & Arun Balaji Buduru. 2019. "A Survey of Black-Box Adversarial Attacks on Computer Vision Models." arXiv:1912.01667.

Birch, Gabriel C., John C. Griffin & Matthew K. Erdman. 2015. *UAS Detection, Classification, and Neutralization: Market Survey 2015*. Albuquerque, NM: Sandia National Laboratories.

Boulanin, Vincent. 2015. "Implementing Article 36 Weapons Reviews in the Light of Increasing Autonomy in Weapon Systems." *SIPRI Insights on Peace and Security*, No. 2015/1. As of 30 November 2020: https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf

——. 2019. 'Artificial Intelligence: A Primer' and 'Promises and perils of artificial intelligence for strategic stability and nuclear risk management: Euro-Atlantic perspectives.' In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. 1, *Euro-Atlantic Perspectives*, edited by Vincent Boulanin, 13–25 and 131-138. 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.

Bulusu, Saikiran, Bhavya Kailkhura, Bo Li, Pramod K. Varshney, & Dawn Song. 2020. "Anomalous Example Detection in Deep Learning: A Survey." *IEEE Access* 8: 132330–47. doi:10.1109/ACCESS.2020.3010274

Buolamwini, Joy, & Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.

Cao, Yulong, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu & Zhuoqing Morley Mao. 2019. "Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, November 2019: 2267–81. doi:10.1145/3319535.3339815

Carvin, Stephanie. 2017. "Normal Autonomous Accidents: What Happens When Killer Robots Fail?" SSRN, 29 April 2018. dx.doi.org/10.2139/ssrn.3161446

Chahal, Husanjot, Ryan Fedasiuk & Carrick Flynn. 2020. *Messier than Oil: Assessing Data Advantage in Military AI*. Washington, DC: Center for Security and Emerging Technology.

Chandola, Varun, Arindam Banerjee & Vipin Kumar. 2009. "Anomaly Detection: A Survey." *ACM Computing Surveys*, July 2009: art. 15. doi:10.1145/1541880.1541882

Chengeta, Thompson. 2020. "Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law." *Denver Journal of International Law & Policy* 45 (1): art. 3.

Christie, Edward Hunter. 2020. "Artificial Intelligence at NATO: Dynamic Adoption, Responsible Use." NATO Review, 24 November. As of 18 January 2021: https://www.nato.int/docu/review/articles/2020/11/24/artificial-intelligence-at-nato-dynamic-adoption-responsible-use/index.html

Cohen, Jeremy. 2018. "Sensor Fusion." Towards Data Science, 22 May. As of 18 January 2021: https://towardsdatascience.com/sensor-fusion-90135614fde6

Cole, Chris. 2019. *Accidents Will Happen: A Review of Military Drone Crash Data as the UK Considers Allowing Large Military Drone Flights in Its Airspace*. Oxford: Drone Wars.

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.

Creswell, Antonia, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta & Anil A. Bharath. 2018. "Generative Adversarial Networks: An Overview." *IEEE Signal Processing Magazine* 35 (1): 53–65. doi:10.1109/MSP.2017.2765202

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

Cummings, M.L., 2017. *Artificial Intelligence and the Future of Warfare*. London: Chatham House.

——. 2019. "Lethal Autonomous Weapons: Meaningful Human Control or Meaningful Human Certification?" *IEEE Technology and Society Magazine* 38 (4): 20–26. doi:10.1109/MTS.2019.2948438

——. 2020a. "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings," *AI Magazine*. As of 15 January 2021: https://hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u36/reality%20check%20final_compressed.pdf

——. 2020b. "The AI That Wasn't There: Global Order and the (Mis)Perception of Powerful AI." In *Policy Roundtable: Artificial Intelligence and International Security*. Texas National Security Review, 2 June. As of 11 December 2020: https://tnsr.org/ roundtable/policy-roundtable-artificial-intelligence-and-international-security

——. 2020c. "Opportunities and Challenges of Deploying Autonomous Delivery Vehicles." ITU News, 15 April. As of 15 January 2021: https://news.itu.int/opportunities-and-challenges-of-deploying-autonomous-delivery-vehicles-missy-cummings

Dahlmann, Anja, & Marcel Dickow. 2019. *Preventive Regulation of Autonomous Weapon Systems: Need for Action by Germany at Various Levels*. Berlin: Stiftung Wissenschaft und Politik.

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, & 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

Danzig, Richard. 2018. *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority.* Washington, DC: Center for a New American Security.

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

Defence and Security Accelerator. 2019. "Competition Document: Autonomy in Challenging Environments." Gov.UK. As of 16 January 2021: https://www.gov.uk/government/publications/competition-autonomy-in-challenging-environments/competition-document-autonomy-in-challenging-environments

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

Defense Innovation Marketplace. n.d. "Test & Evaluation, Validation & Verification." As of 18 January 2021: https://defenseinnovationmarketplace.dtic.mil/technology-interchange-meetings/autonomy-tim/autonomy_tevv

Department of the Army. 2014. *Risk Management*. ATP 5-19, C1. Washington, DC.

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.

———.2021. "Normative Epistemology for Lethal Autonomous Weapons Systems." In *Lethal Autonomous Weapons: Re-Examining the Law & Ethics of Robotic Warfare*, edited by J. Galliot, D. MacIntosh & J. D. Ohlin. Oxford University Press.

Devitt, S. Kate, Jason Scholz, Timo Schless & Larry Lewis. 2020. "Developing a Trusted Human-AI Network for Humanitarian Benefit." Draft submitted for peer review to *Digital War*.

Dickson, Ben. 2020. "All You Need to Know about Symbolic Artificial Intelligence." The Next Web, May 2020. As of 4 January 2021: https://thenextweb.com/neural-basics/2020/06/02/all-you-need-to-know-about-symbolic-artificial-intelligence

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

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.

Elish, Madeleine Claire. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5. doi: 10.17351/ests2019.260

European Cockpit Association (ECA). 2019. "Specific Operations Risk Assessment (SORA)." 28 January. As of 4 March 2021: https://www.eurocockpit.be/positions-publications/specific-operations-risk-assessment-sora

European Commission. 2020. *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*, EU document COM(2020) 64 final, 19 February 2020.

European Union Agency for Fundamental Rights (FRA). 2019. *Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights*. Vienna.

Evas, Tatjana. 2018. *A Common EU Approach to Liability Rules and Insurance for Connected and Autonomous Vehicles*. PE 615.635. Brussels: European Parliamentary Research Service.

Farrant, James, & Christopher M. Ford. 2017. "Autonomous Weapons and Weapon Reviews: The UK Second International Weapon Review Forum." *International Law Studies* 93 (389): 390–422.

Flournoy, Michèle A., Avril Haines & Gabrielle Chefitz. 2020. *Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems*. Washington, DC: WestExec Advisors.

Freedberg, Sydney J. 2018. "Big Bad Data: Achilles' Heel of Artificial Intelligence." Breaking Defense, 13 November, 4.24 p.m. As of 2 December 2020: https://breakingdefense.com/2018/11/big-bad-data-achilles-heel-of-artificial-intelligence

———. 2019. “Pentagon’s AI Problem Is ‘Dirty’ Data: Lt. Gen. Shanahan.” Breaking Defense, 13 November, 9.52 a.m. As of 18 January 2021: https://breakingdefense.com/2019/11/exclusive-pentagons-ai-problem-is-dirty-data-lt-gen-shanahan

———. 2020. “AI’s Data Hunger Will Drive Intelligence Collection.” Breaking Defense, 3 September, 1.55 p.m. As of 18 January 2021: https://breakingdefense.com/2020/09/ais-data-hunger-will-drive-intelligence-collection

French, Paddy, Gijs Krijnen & Fred Roozeboom. 2016. “Precision in Harsh Environments.” *Microsystems & Nanoengineering*, 2: 16048. doi:10.1038/micronano.2016.48

Gates, Dominic, & Mike Baker. 2019. “The Inside Story of MCAS: How Boeing’s 737 MAX System Gained Power and Lost Safeguards.” *Seattle Times*, 22 June, 2.00 p.m. As of 16 January 2021: https://www.seattletimes.com/seattle-news/times-watchdog/the-inside-story-of-mcas-how-boeings-737-max-system-gained-power-and-lost-safeguards

Geiß, Robin. 2017. “Autonomous Weapons Systems: Risk Management and State Responsibility.” In *Lethal Autonomous Weapons Systems: Technology, Definition, Ethics, Law & Security* edited by Robin Geiß, 109-118. Federal Foreign Office.

Gershgorn, Dave. 2017. “The Data That Transformed AI Research – And Possibly the World.” Quartz, 26 July. As of 25 January 2021: https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world

Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford & Chris Olah. 2021. “Multimodal Neurons in Artificial Neural Networks.” *Distill* 6: 3. doi: 10.23915/distill.00030

Goldblum, Micah, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li & Tom Goldstein. 2020. “Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses.” arXiv:2012.10544

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

Grother, Patrick J., Mei L. Ngan & Kayee K. Hanaoka. 2019. *Face Recognition Vendor Test Part 3: Demographic Effects*. Washington, DC: National Institute of Standards and Technology.

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.

———. 2020. *2020 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: Chairperson’s Summary*. (Advance Copy).

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, 32-38. Stockholm: Stockholm International Peace Research Institute.

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.

Heaven, Will Douglas. 2020. “Our Weird Behavior during the Pandemic Is Messing with AI Models.”

MIT Technology Review, 11 May 2021. 19 November 2020: https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop

Hellström, Tomas, Virginia Dignum, & Suna Bensch. 2020. "Bias in Machine Learning - What Is It Good For?" arXiv:2004.00686

Herpig, Sven. 2019. *Securing Artificial Intelligence*. Berlin: Stiftung Neue Verantwortung.

Heyns, Christof. 2014. "Speaker's Summary: Increasingly Autonomous Weapon Systems: Accountability and Responsibility." In *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. Geneva: International Committee of the Red Cross.

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.

Holland Michel, Arthur. 2020a. *Counter Drone Systems, 2nd Edition*. Annandale-on-Hudson, NY: Center for the Study of the Drone.

———. 2020b. *The Black Box, Unlocked: Predictability and Understandability in Military AI.* Geneva: United Nations Institute for Disarmament Research. doi.org/10.37559/SecTec/20/AI1

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

Ibarz, Julian, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor & Sergey Levine. 2021. "How to Train Your Robot with Deep Reinforcement Learning; Lessons We've Learned." *Journal of Robotics Research*. doi:10.1177/0278364920987859

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

Institute of Electrical and Electronics Engineers (IEEE). 2017. "Reframing Autonomous Weapons Systems." As of 25 January 2021: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_reframing_autonomous_weapons_v2.pdf

Institute of Electrical and Electronics Engineers Standards Association (IEEE SA). 2019. "IEEE/ISO/IEC 15026-1_Revision-2019 – ISO/IEC/IEEE International Standard – Systems and software engineering – Systems and software assurance – Part 1: Concepts and vocabulary." As of 5 March 2021: https://standards.ieee.org/standard/15026-1_Revision-2019.html

International Committee of the Red Cross (ICRC). 2006. "A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977." *International Review of the Red Cross* 864: 931–56

———. 2014. *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. Geneva.

———. 2018. *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* Geneva.

———. 2019. "Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach." ICRC, 6 June. As of 8 January 2021: https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach

Jafarnia-Jahromi, Ali, Ali Broumandan, John Nielsen, & Gérard Lachapelle. 2012. "GPS Vulnerability to Spoofing Threats and a Review of Antispoofing Techniques." *International Journal of Navigation and Observation*. doi:10.1155/2012/127072

Jain, Ashesh, Lei Zhang & Li Jiang. 2019. "High-Fidelity Sensor Calibration for Autonomous Vehicles." Lyft Self-Driving, 14 August. As of 11 January 2021: https://medium.com/lyftself-driving/high-fidelity-sensor-calibration-for-autonomous-vehicles-6af06eba4c26

Jenihhin, Maksim, Matteo Sonza Reorda, Aneesh Balakrishnan & Dan Alexandrescu. 2019. "Challenges of Reliability Assessment and Enhancement in Autonomous Systems." *32nd IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*. arXiv:1909.03040

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

Kantchelian, Alex, Sadia Afroz, Ling Huang, Aylin Caliskan Islam, Brad Miller, Michael Carl Tschantz, Rachel Greenstadt, Anthony D. Joseph & J.D. Tygar. 2013. "Approaches to Adversarial Drift." In Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security: 99–110. doi:10.1145/2517312.2517320

Kerns, Andrew J., Daniel P. Shepard, Jahshan A. Bhatti & Todd E. Humphreys. 2014. "Unmanned Aircraft Capture and Control via GPS Spoofing." *Journal of Field Robotics* 31 (4): 617-36. doi:10.1002/rob.21513

Kim, Seong Dae. 2012. "Characterizing Unknown Unknowns," paper, *PMI Global Congress 2012 – North America*, Vancouver, 20–23 October 2012. As of 2 March 2021: https://www.pmi.org/learning/library/characterizing-unknown-unknowns-6077

Koopman, Philip, & Frank Fratrik. 2019. "How Many Operational Design Domains, Objects, and Events?" preprint, *Safe AI 2019: AAAI Workshop on Artificial Intelligence Safety*, Honolulu, 27 January. As of 18 January 2021: https://users.ece.cmu.edu/~koopman/pubs/Koopman19_SAFE_AI_ODD_OEDR.pdf

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

Kouw, Wouter M., & Marco Loog. 2018. "An Introduction to Domain Adaptation and Transfer Learning." arXiv:1812.11806

Krishnamurthy, Prabhakar. 2019. "Understanding Data Bias." Towards data science, 12 September. As of 16 January 2021: https://towardsdatascience.com/survey-d4f168791e57

Leveson, Nancy. 1995. *SafeWare: System Safety and Computers*. Boston: Addison-Wesley.

Lewis, Dustin. 2019. "Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider." Humanitarian Law & Policy, 21 March. As of 18 January 2021: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider

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.

Li, Jessie. 2019. "Fooling Facial Recognition with Fashion." Axios, 7 September. As of 15 January 2021: https://www.axios.com/fooling-facial-recognition-fashion-06b04639-7e47-4b55-aa00-82410892a663.html

Lipton, Zachary. 2020. Remarks at *The Technical Limits of AI Ethics*, New York, 16 December. As of 25 January 2021: https://www.carnegiecouncil.org/studio/multimedia/20201217-technical-limits-artificial-intelligence

Llorens, Ashley. 2020. Remarks at *Building Trust Through Testing*, Washington, DC, 6 October. As of 15 January 2021: https://cset.georgetown.edu/event/building-trust-through-testing

Lohn, Andrew J. 2020a. "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance." arXiv:2009.00802

——. 2020b. *Attacking AI: A Primer for Policymakers on Machine Learning Cybersecurity*. Washington, DC: Center for Security and Emerging Technology.

Lubell, Noam, Jelena Pejic & Claire Simmons. 2019. *Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy, and Good Practice*. Geneva: International Committee of the Red Cross.

Luckcuck, Matt, Marie Farrell, Louise A. Dennis, Clare Dixon & Michael Fisher. 2019. "Formal Specification and Verification of Autonomous Robotic Systems: A Survey." *ACM Computing Surveys* 52 (5). doi:10.1145/3342355

Maas, Matthijs M. 2018. "Regulating for 'Normal AI Accidents' – Operational Lessons for the Responsible Governance of AI Deployment." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 223–28. doi:10.1145/3278721.3278766

"Machine Learning Glossary." n.d. Google Developers. As of 16 January 2021: https://developers.google.com/machine-learning/glossary

Mann, Jindan-Karena. 2019. "Autonomous Weapons Systems and the Liability Gap, Part One: Introduction to Autonomous Weapons Systems and International Criminal Liability." Rethinking SLIC, 15 July. As of 8 December 2020: https://www.rethinkingslic.org/blog/criminal-law/51-autonomous-weapons-systems-and-the-liability-gap-part-one-introduction-to-autonomous-weapon-systems-and-international-criminal-liability

Margulies, Peter. 2019. "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. As of 14 April 2021: http://ssrn.com/abstract=2734900

Meyer-Vitali, André, Roos Bakker, Michael van Bekkum, Maaike de Boer, Gertjan Burghouts, Jurriaan van Diggelen, Judith Dijk, Corrado Grappiolo, Joachim de Greeff, Albert Huizing & Stephan Raaijmakers. 2019. *Hybrid AI: White Paper*. The Hague: Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek.

Milani, Peter A. 2020. "Autonomous Weapon Systems for the Land Domain." The Cove, 8 October. As of 15 January 2021: https://cove.army.gov.au/article/autonomous-weapon-systems-the-land-domain

Military Aviation Authority (MAA). 2014. *Manual of Air Safety (MAS)*. London: Defence Safety Authority. As of 4 March 2021: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/912759/MAS_Issue_7.pdf

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.

Mulchandani, Nand. 2020. Remarks at *Virtual Event and Live Q&A with Mr. Nand Mulchandani, Joint Artificial Intelligence Center*, 25 August. As of 18 January 2021: https://securityandtechnology.org/virtual-library/video/virtual-event-and-live-qa-with-mr-nand-mulchandani-acting-director-of-the-u-s-department-of-defense-joint-artificial-intelligence-center

Narayan Banerjee, Debarag, & Sasanka Sekhar Chanda. 2020. "AI Failures: A Review of Underlying Issues." arXiv:2008.04073

National Coordination Office for Space-Based Positioning, Navigation, and Timing. 2021. "GPS Accuracy." GPS.gov. As of 4 January: https://www.gps.gov/systems/gps/performance/accuracy/#speed

National Security Commission on Artificial Intelligence (NSCAI). 2021. *Final Report*. Washington, DC.

Naval Studies Board. 2005. *Autonomous Vehicles in Support of Naval Operations*. Washington, DC: National Academies Press.

Nelson, Kevin, George Corbin & Misty Blowers. 2014. "Evaluating Data Distribution and Drift Vulnerabilities of Machine-Learning Algorithms in Secure and Adversarial Environments." In *Proceedings Volume 9119, Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII*. doi:10.1117/12.2053045

Nguyen, Ahn, Jason Yosinski & Jeff Clune. 2015. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," paper, *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7–12 June 2015. doi:10.1109/CVPR.2015.7298640

Nohle, Ellen, & Isabel Robinson. 2017. "War in Cities: The 'Reverberating Effects' of Explosive Weapons." Humanitarian Law and Policy, 2 March. As of 18 January 2021: https://blogs.icrc.org/law-and-policy/2017/03/02/war-in-cities-the-reverberating-effects-of-explosive-weapons

North Atlantic Treaty Organization (NATO). 2016. *NATO Standard AJP-3.9: Allied Joint Doctrine for Joint Targeting* A(1). Brussels.

OpenAI. 2017. "Attacking Machine Learning with Adversarial Examples." As of 11 January 2021: https://openai.com/blog/adversarial-example-research

Pan, Jiyan, & Bo Hu. 2007. "Robust Occlusion Handling in Object Tracking," paper, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, 17–22 June 2007. doi:10.1109/CVPR.2007.383453

Parasuraman, Raja, & Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52 (3): 381–410. doi:10.1177/0018720810376055

Perry, Tal. 2019. "Four Mistakes You Make When Labeling Data." Towards data science, 28 May. As of 2 December 2020: https://towardsdatascience.com/four-mistakes-you-make-when-labeling-data-7e431c4438a2

Pinelis, Jane. 2020. Speaking at *Webinar Series on the Technological, Military and Legal Aspects of Lethal Autonomous Weapon Systems – Technological Aspects*, 26 October. As of 6 April 2021: https://www.unidir.org/events/webinar-series-technological-military-and-legal-aspects-lethal-autonomous-weapon-systems

Pokhrel, Sabina. 2020. "6 Obstacles to Robust Object Detection." Towards data science, 11 February. As of 16 January 2021: https://towardsdatascience.com/6-obstacles-to-robust-object-detection-6802140302ef

Pontin, Jason. 2018. "Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning." *Wired*, 2 February. As of 16 January 2021: https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning

Quéguiner, Jean-François. 2006. "Precautions under the Law Governing the Conduct of Hostilities." *International Review of the Red Cross* 88 (864), December 2006: 793–821.

Rabanser, Stephan, Stephan Gunnemann & Zachary C. Lipton. 2019. "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift," paper, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, 8–14 December 2019. arXiv:1810.11953

Rajaei, Karim, Yalda Mohsenzadeh, Reza Ebrahimpour & Seyed-Mahdi Khaligh-Razavi. 2019. "Beyond Core Object Recognition: Recurrent Processes Account for Object Recognition under Occlusion." *PLoS Computational Biology* 15. doi:10.1371/journal.pcbi.1007001

Ramisa Ayats, Arnau, David Aldavert, Shrihari Vasudevan, Ricardo Toledo & Ramon Lopez de Mantaras.

2012. “Evaluation of Three Vision Based Object Perception Methods for a Mobile Robot.” *Journal of Intelligent and Robotic Systems* 68 (2): 185–208. doi:10.1007/s10846-012-9675-8

Reim, Garrett. 2020. “US Air Force Grapples with Vexing Problem of AI Spoofing.” FlightGlobal, 1 September. As of 18 January 2021: https://www.flightglobal.com/defence/us-air-force-grapples-with-vexing-problem-of-ai-spoofing/139973.article

Ren, Jie, & Balaji Lakshminarayanan. 2019. “Improving Out-of-Distribution Detection in Machine Learning Models.” Google AI Blog, 17 December. As of 18 January 2021: https://ai.googleblog.com/2019/12/improving-out-of-distribution-detection.html

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

Sassóli, Marco. 2014. “Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified.” *International Law Studies* 90: 308–40.

Scharre, Paul. 2016. *Autonomous Weapons and Operational Risk*. Washington, DC: Center for a New American Security. As of 15 January 2021: https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk

——. 2019. “Military Applications of Artificial Intelligence: Potential Risks to International Peace and Security.” In *The Militarization of Artificial Intelligence*. New York: Stanley Center for Peace and 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

Schmidt, Eric. 2021. Remarks at *NSCAI Plenary*, 25 January. As of 28 January 2021: https://www.youtube.com/watch?v=OxM8pC_mjCI

Schmitt, Michael, & Lt. Col. Matthew King. 2018. “The ‘Shift Cold’ Military Tactic and International Humanitarian Law.” Just Security, 20 February. As of 1 December 2020: https://www.justsecurity.org/52198/shift-cold-tactic-international-humanitarian-law

Schwarz, Elke. 2018. “The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems.” Humanitarian Law and Policy, 29 August. As of 3 December 2020: https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems

Scott, Peter J., & Roman V. Yampolskiy. 2019. “Classification Schemas for Artificial Intelligence Failures.” arXiv:1907.07771

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary & Michael Young. 2014. “Machine Learning: The High Interest Credit Card of Technical Debt,” paper, *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. As of 16 January 2021: https://research.google/pubs/pub43146

Shakhuro, Vlad, Boris Faizov & Anton S. Konushin. 2019. “Rare Traffic Sign Recognition Using Synthetic Training Data.” In *Proceedings of the 3rd International Conference on Video and Image Processing*, 23–26. doi:10.1145/3376067.3376105

Shendre, Sushrut. 2020. “Model Drift in Machine Learning.” Towards data science, 13 May. As of 16 January 2021: https://towardsdatascience.com/model-drift-in-machine-learning-models-8f7e7413b563

Shermeyer, Jake. 2019. “RarePlanes - An Introduction.” The DownLinQ, 10 December. As of 18 January 2021: https://medium.com/the-downlinq/rareplanes-an-introduction-b28449222ca4

Smith, Kevin A., Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Joshua B. Tenenbaum & Tomer D. Ullman. 2019. "Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations," paper, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. As of 6 April 2021: http://physadept.csail.mit.edu/papers/adept.pdf

Sokol, Joshua. 2018. "Why Artificial Intelligence Like AlphaZero Has Trouble with the Real World." *Quanta Magazine*, 21 February. As of 16 January 2021: https://www.quantamagazine.org/why-alphazeros-artificial-intelligence-has-trouble-with-the-real-world-20180221

Strout, Nathan. 2020. "Could This Software Help Users Trust Machine Learning Decisions?" Defense-News, 20 July. As of 18 January 2021: https://www.defensenews.com/artificial-intelligence/2020/07/20/could-this-software-help-users-trust-machine-learning-decisions

Stumborg, Michael. 2019. "See you in a Month: AI's Long Data Trail." War on the Rocks, 17 October. As of 18 January 2021: https://warontherocks.com/2019/10/see-you-in-a-month-ais-long-data-tail

Svenmarck, Peter, Linus Luotsinen, Mattias Nilsson & Johan Schubert. 2018. *Possibilities and Challenges for Artificial Intelligence in Military Applications*. Brussels: NATO Science & Technology Organization.

Taori, Rohan, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht & Ludwig Schmidt. 2020. "Measuring Robustness to Natural Distribution Shifts in Image Classification," paper, *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. As of 6 April 2021: https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

Tholl, Hans Dieter. 2018. "Review and Prospects of Optical Countermeasure Technologies." In *Proceedings Volume 10797, Technologies for Optical Countermeasures XV*. doi:10.1117/12.2326594

Tremblay, Jonathan, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon & Stan Birchfield. 2018. "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization," paper, *CVPR 2018 Workshop on Autonomous Driving*, Salt Lake City, 18–22 June 2018. arXiv:1804.06516

Tucker, Patrick. 2019. "The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth." Defense One, 31 March. As of 18 January 2021: https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944

United Nations Institute for Disarmament Research (UNIDIR). 2016. *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*. Geneva.

———. 2018. *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies*. Geneva.

United States Air Force (USAF). 2019. *Annex 3-60 Targeting*.

Verdiesen, Ilse, Filippo Santoni de Sio & Virginia Dignum. 2021. "Accountability and Control over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight." *Minds and Machines* 31: 137–63. doi:10.1007/s11023-020-09532-9

Wand, Yair, & Richard Y. Wang. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM*, November. doi:10.1145/240455.240479

Wang, Angtian, Yihong Sun, Adam Kortylewski & Alan Yuille. "Robust Object Detection under Occlusion with Context-Aware CompositionalNets," In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12642–51. New York: IEEE. doi:10.1109/CVPR42600.2020.01266

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

Weir, Nick. 2018. "Challenges with SpaceNet 4 Off-Nadir Satellite Imagery: Look Angle and Target Azimuth Angle." The DownLinQ, 9 November. As of 10 May 2021: https://medium.com/the-downlinq/challenges-with-spacenet-4-off-nadir-satellite-imagery-look-angle-and-target-azimuth-angle-2402bc4c3cf6

Yampolskiy, Roman V. 2020. "AI Will Fail, Like Everything Else, Eventually." Mind Matters, 14 July. As of 3 December 2020: https://mindmatters.ai/2020/07/ai-will-fail-like-everything-else-eventually

Yampolskiy, Roman V., & M.S. Spellchecker [:)]. 2016. "Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures." arXiv:1610.07997

Yeh, Enoch R., Junil Choi, Nuria G. Prelcic, Chandra R. Bhat & Robert W. Heath, Jr. 2017. "Security in Automotive Radar and Vehicular Networks." *Microwave Journal* 60 (5).

Yurtsever, Ekim, Jacob Lambert, Alexander Carballo & Kazuya Takeda. 2020. "A Survey of Autonomous Driving: Common Practices and Emerging Technologies." *IEEE Access* 8: 58443–69. doi:10.1109/ACCESS.2020.2983149

KNOWN UNKNOWNS

Data Issues and Military Autonomous Systems

In order to perform as desired, autonomous systems must collect data that are complete, relevant, accurate, and aligned with the data for which the system was developed and tested. But the harsh, dynamic, complex and adversarial nature of conflict environments poses a wide range of obstacles to the collection of such data. As a result, autonomous systems cannot always be expected to achieve the exact same performance in real-world use that they demonstrated in development or testing. And crucially, they will be liable to failures that are both inevitable and impossible to anticipate: "known unknowns." This report describes common data issues for autonomous systems and explains how they give rise to "known unknown" failures. It then explores the legal and operational implications of such failures, and considers a range of potential policy and technical solutions by which they could be addressed.

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH