Digital Society (2022) 1:11
https://doi.org/10.1007/s44206-022-00014-2

ORIGINAL PAPER

# The Challenge of Ethical Interoperability

**David Danks**[1] · **Daniel Trusilo**[2]

Received: 26 February 2022 / Accepted: 26 July 2022 / Published online: 10 August 2022


## Abstract

Defense organizations are increasingly developing ethical principles to guide the design, development, and use of responsible AI, most notably for defense, security, and intelligence uses. While these principles have the potential to lead to more ethical and responsible uses of AI, they also create a novel issue that we term the *challenge of ethical interoperability*. In particular, we start with the observation that the frequent necessity for collaboration between defense organizations results in their possible need to use one another's AI systems, either directly or indirectly. In such cases, the AI system presumably satisfies the originator organization's ethical principles, but the adopting organization should only use it if it also satisfies their AI principles. One might naturally consider using the operating characteristics of the AI system to establish ethical interoperability, as those parallel the features used for technical interoperability. However, we argue that if the operating characteristics are sufficient to enable such assurance, then they will be too detailed to be disclosed. Instead, we propose a system of self-certification that provides adopting organizations with assurances that a system satisfies their ethical AI principles. We argue that our proposed framework properly aligns incentives to ensure ethical interoperability between defense organizations. Moreover, although this challenge is particularly salient for defense organizations, ethical interoperability must also be addressed in non-defense settings by organizations that have developed ethical AI principles.

**Keywords** Ethical AI · Responsible AI · Ethical principles · Ethical interoperability · Defense AI

David Danks and Daniel Trusilo contributed to all aspects of this manuscript.

✉ David Danks
ddanks@ucsd.edu

[1] Halicioğlu Data Science Institute & Department of Philosophy, University of California, San Diego, La Jolla, CA, USA

[2] School of Economics and Political Science, University of St. Gallen, St. Gallen, Switzerland

## 1 The Need for Ethical Interoperability

Recent headlines have been filled with examples of significant harms resulting from the use of AI systems across a wide range of domains. One reaction has been repeated calls for organizations to adopt principles and practices that, in theory, help to ensure that AI systems are designed, developed, deployed, and used in responsible and ethical ways. These calls have been particularly widespread for AI uses in high-consequence settings and domains, including defense and security. Principles, frameworks, and methods for ethical and responsible use of AI have increasingly been adopted by governments and national defense organizations, as well as related private companies and nonprofit institutions. Substantial efforts are now being devoted to translating those principles into appropriate processes, training, and policies for the acquisition, testing, evaluation, and deployment of AI systems.

These principles are generally laudable, but their diversity raises a new problem for governance of AI systems in defense and security contexts. In particular, similar-but-different principles are being adopted by different parties. When those groups subsequently want or need to collaborate, they each must determine whether others' AI systems conform to their own ethical principles. That is, these organizations need to determine whether some AI system is *ethically interoperable* in the sense that the AI is consistent with the adopting organization's principles and systems. For example, suppose nations *A* and *B* are engaged in a combined peacekeeping operation, and *A* has an AI-augmented surveillance system that conforms to *A*'s ethical principles. That system should only be used as part of the combined operation if it *also* conforms to *B*'s ethical principles, just as a radio system should only be used in combined operations if it is operable with both nations' communication networks. Alternatively, suppose that nations *C* and *D* want to cooperate in pursuing a terrorist network, and *C* uses a semi-autonomous natural language processing (NLP) system to filter electronic communications. Nation *D* should (ethically) only cooperate if they can determine that this system, and the resulting intelligence, follows *D*'s principles for responsible use of technology.

There are multiple types of interoperability, including technical integration, shared codes of conduct or behavior, legal commitments, and common tactics. In particular, AI systems can present significant challenges and opportunities around technical interoperability (Reeves, 2021). However, technical and ethical interoperability are conceptually independent and should be considered separately. Our focus here is not when different AI systems *could* (technically) be integrated but when they *should* (ethically) be integrated. Moreover, any answer to this question must provide practical, operational guidance (not mere words). There are important proposals and debates about what ethical AI principles ought to be adopted by defense organizations (e.g., Taddeo et al., 2021), but we focus on how those organizations can operationalize and live up to their ethical AI principles, whatever they might be.

One might hope that the problem of ethical interoperability would be only a theoretical issue; perhaps it never actually arises or is easily solved in real-world

contexts. Unfortunately, that hope is in vain: Sect. 1.1 demonstrates that ethical interoperability is potentially a serious practical challenge in ethical governance of AI systems in defense, security, and intelligence contexts. Section 2 then considers several natural approaches to solving this challenge, including through identification of the ranges of operating characteristics that satisfy a set of ethical principles. Given such mappings, interoperability could be established by showing that the AI has acceptable operating characteristics from the perspective of both sets of principles. Unfortunately, we argue that such mappings are possible only if we construe "operating characteristics" very broadly, and almost certainly more broadly than any defense organization would find acceptable. These potential paths to ethical interoperability would require far more transparency and disclosure than is reasonable to expect in the domains of defense, security, and intelligence.

Section 3 presents an alternative approach that starts with the now-common observation that there is significant overlap in the practical content of various sets of ethical AI principles despite differences in terminology (Fjeld et al., 2020). This overlap in content leads to overlap in permissible AI systems and uses. For example, the ethical AI principles of essentially every national defense organization allow for the use of (well-designed) AI systems to optimize vehicle maintenance. Ethical interoperability should not be an issue for appropriate uses of such maintenance optimization systems (i.e., an adopter organization could ethically use the AI). We contend that the substantive operational cores shared between various sets of principles can provide the foundation for a system of self-certification that will enable ethical interoperability in many, though not all, circumstances. There is unlikely to be a domain- or system-general solution to the challenge of ethical interoperability, but our proposal will likely enable defense organizations to continue much of their current cooperation while expanding shared AI use in ethical ways. Of course, the challenge of ethical interoperability almost certainly applies in many non-defense settings, and so Sect. 4 explores other instances of the challenge and how our proposed resolution could apply. Section 5 presents opportunities for future research and our conclusion. But first, we consider whether ethical interoperability is a real-world concern or a mere theoretical possibility.

### 1.1 Ethical Interoperability in Practice

Combined operations typically involve defense elements from different nations working collaboratively towards a common goal. Such multinational efforts thus provide obvious situations in which ethical interoperability might be necessary. In fact, the North Atlantic Treaty Organization (NATO) has developed a strategy (NATO, 2021) that explicitly mentions the importance of interoperability of responsible AI, though the focus is predominantly on technical issues rather than ethical ones. We discuss the NATO effort further in Sect. 2, but it is essential to note from the outset that this agreement indicates awareness of distinctive challenges of AI deployed collaboratively and across organizational boundaries.

At the same time, resolution of those challenges is complicated by the fact that, at the time of writing, there are few sets of ethical AI principles that have officially been adopted by a national defense organization (Stanley-Lockman, 2021). The most prominent set is the US Department of Defense (DoD) *Ethical Principles for AI,* adopted in February 2020 (DoD, 2020), based on an earlier proposal by the Defense Innovation Board (DIB, 2019). Very recently, the Defence AI Strategy 2022 of the UK Ministry of Defence (MoD, 2022b) included a set of principles for the "ambitious, safe, responsible" use of defense AI (MoD, 2022a). Other nations have signaled that their defense organizations will adopt official ethical AI principles in the near future. Our discussion of ethical interoperability in the multinational context will thus rely partly on official government documents that articulate national positions on the ethical use of AI for defense, even when those might not necessarily have binding force. We suggest that these documents clearly signal the ethical principles guiding decision-making around AI by that nation's defense organization. Importantly, these signals indicate significant differences between different defense organizations, whether in the actual sets of principles or the (implied) ways to balance principles if they conflict (Whittlestone et al., 2019).

In particular, several allies are pursuing national strategies around (ethical) AI that involve notions of sovereignty. For example, the 2019 French AI Task Force document *Artificial Intelligence in Support of Defence* links the principles of Trustworthy and Responsible AI inextricably to the French government's ability to audit both algorithms and the data used to train them. The document explicitly clarifies that "preserving digital sovereignty therefore also involves controlling the algorithms and their configuration, and the governance of data" (AITF, 2019). Similarly, Australia includes the notion of sovereignty as part of its discussion of ethical AI for defense. The Australian Defence Science and Technology Group's *A Method for Ethical AI in Defence* (hereafter Method; Devitt et al., 2020) lists five facets of Australia's approach to ethical AI for defense: responsibility, governance, trust, law, and, traceability. Under the facet of trust, the method highlights sovereign capability as a concern: "lack of investment in sovereign AI could impact [Australia's] ability to achieve sovereign decision superiority" (Devitt et al., 2020, p. 21). As Devitt and Copeland (2021) further clarify: "To be trusted AI systems need to be safe and secure within the Nation's sovereign supply chain" (p. 22).

Ethical principles based in national sovereignty present an obvious problem for ethical interoperability, as it is difficult to imagine a situation in which an originator country would cede ownership of its AI algorithm. Defense AI systems can be among a nation's most treasured assets, and it is improbable that any country would cede control or ownership to another nation, even in a combined operation involving allies. However, sovereignty principles imply that the adopter country can only ethically use the AI system (or its products) if they have such control. If a country insists on sovereign control as part of its ethical AI principles or plan, interoperability with other nations will almost surely be challenging (even if such integration might be technically feasible). Therefore, a method of addressing different AI principles, even among allies, is essential.

As a concrete example, the Royal Australian Air Force's (RAAF) Plan Jericho aims to develop an integrated, fifth-generation force, using AI to process and integrate vast

amounts of information from advanced sensor platforms at machine speed (RAAF, 2019). At the same time, the US Air Force and Space Force are developing the Advanced Battle Management System (ABMS) to disaggregate sensors and command and control systems as part of DoD's Combined Joint All-Domain Command and Control effort (Hoehn, 2022a, b). The ABMS, like the RAAF's Plan Jericho, relies heavily on the use of AI to process and share data across services and allow faster decision-making.

In coalition operations, the US DoD assets will presumably use AI to process and incorporate data from satellite imagery as part of the ABMS, and then want to share those data and recommendations with RAAF assets. Such exchange depends on ethical interoperability: do the data and recommendations generated by the US AI systems conform to the *Australian* ethical principles? Since sovereignty is a key part of those latter principles, particularly the principle of trust, the exchange of data and AI recommendations would (in theory) require RAAF assets to have significant control over the ABMS (which would presumably never be allowed by the US). Even if national sovereignty required "only" detailed knowledge of the algorithms and training data underlying the AI, it is unlikely that the US DoD would disclose that information, even to facilitate coalition air operations. Much of the focus and effort in establishing coalition operations has traditionally been on technical issues, such as ensuring that RAAF elements can technologically plug into the US DoD network. However, interoperability is not simply a technical matter; it also requires careful attention to the ethical principles of all parties.

One might look at this example and think that national boundaries are the principal barrier to interoperability. However, the challenge of ethical interoperability can also arise between agencies within the same country. For example, one of the five DoD principles is that AI systems must be Governable, which includes the requirement that a human can disengage or deactivate an operationally deployed AI system that behaves in unintended ways. In contrast, the *Principles of AI Ethics for the Intelligence Community* (IC) provided by the US Office of the Director of National Intelligence (ODNI) make no reference to governability (ODNI, 2020b). ODNI's *AI Ethics Framework for the IC* expands on its Principles by outlining their implementation (ODNI, 2020a), and that document does emphasize the importance of humans being involved in decision processes depending on the assessed risk level of specific AI systems. This emphasis might appear to be similar to the DoD's principle of governability but actually derives from ODNI's ethical principles of Responsibility and Accountability. Similarly, the ODNI framework states that agencies should identify individuals who have the authority to modify, limit, or stop a system in response to concerns (ODNI, 2020a). However, this emphasis again stems from ODNI's focus on responsibility and accountability rather than governability.

Now consider an AI system operated by the IC to process geospatial imagery and find specific information, objects, or events in those vast datasets. Project Convergence in 2020 linked such systems to DoD artillery units, enabling data from IC-operated satellites to be used for targeting by US Army artillery in less than twenty seconds (Feickert, 2021; Freedberg, 2020). More specifically, open-source reports suggest that (a) the IC's AI system was analyzing geospatial imagery to identify target locations; (b) those processed data were sent to a DoD command and control

element; and (c) this element communicated targeting data to a US Army Fire Direction Center. From the IC perspective, there is no ethical need to be able to disconnect or deactivate the AI system; responsibility and accountability require only that the IC determine the reliability of the system's analyses, particularly since (from their perspective) no human life is directly at stake. From the US Army's perspective, they are using an AI system (albeit indirectly), so it must be governable: that is, the AI image-processing system must be capable of being deactivated, or at least disconnected from the targeting decisions, if it exhibits unintended behaviors. We thus see a concrete example in which an AI system satisfies the originator's ethical AI principles but not the adopter's ethical AI principles. That is, we contend that this example potentially represents a failure of *ethical* interoperability, despite the successful *technical* interoperability.[1] Moreover, the challenge will likely only worsen in the future as Project Convergence aims to include other nations (Feickert, 2021).

## 2 Interoperability via Operating Characteristics?

Ethical interoperability is a significant challenge for the governance of AI systems in modern defense organizations, given the widespread formal and informal cooperation that occurs between nations and between government and private industry. Perhaps the most obvious path towards addressing this problem would focus on finding some shared or common terminology in which the various principles could be expressed. In fact, the recently adopted NATO Principles of Responsible Use for AI in Defence (NATO, 2021) seem to provide such terminology: they are "Based on existing and widely accepted ethical, legal, and policy commitments" and "are aimed at providing coherence for both NATO and Allies to enable interoperability." However, the document also explicitly states that "These Principles do not affect or supersede existing obligations and commitments, both national and international," and so they do not solve the challenge of ethical interoperability. If a country's principles require sovereignty, for example, then the issues noted in Sect. 1.1 will still arise since the sovereignty requirement will supersede the NATO principles. More generally, the NATO principles provide only a minimum baseline for ethically responsible AI for allied countries. An originator nation's AI system could, in theory, satisfy the NATO principles without fully satisfying a NATO ally adopter nation's national-level principles. Perhaps most importantly, the NATO principles do not provide operationalizations that can be used by adopter organizations to verify or confirm ethical interoperability but instead call themselves for the creation of such operationalizations, including "best practices for Allies." Our focal question

[1] In fairness, it is unreasonable to have expected ethical interoperability here, given that the DoD principles were approved 6 months prior to the first Project Convergence test, and the ODNI principles only 1 month prior to the test. In addition, it is possible that the IC assets actually can be deactivated in the ways that the DoD principle of Governability requires. Nonetheless, this example reveals the ways that ethical interoperability is potentially an issue in everyday joint operations.

here is precisely how to operationalize this need and therefore complements the NATO agreement.

More generally, one does not find the required shared language for substantive normative ethical theories, even outside of the defense sector: terms such as "privacy" or "security" do not have universally accepted meanings, nor is there a more fine-grained set of terms that could enable the translation of all ethical principles into a single language. Although shared ideas underlie many sets of ethical principles (Fjeld et al., 2020; Floridi & Cowls, 2021), there is unlikely to be a shared language.[2]

Instead, one might aim to develop compatible regulatory regimes in which the defense organizations use similar development, testing, and certification processes. Proposals of "best practices" for ethical and responsible AI (e.g., from the US Defense Innovation Unit (DIU); Dunnmon et al., 2021) arguably increase the likelihood of appropriately similar practices. However, even if the adopter organization knows that the originator uses the same processes to determine ethical compliance, they do not necessarily know that the system is ethically interoperable since the originator's processes would be based on the originator's principles and values, not those of the adopter.

The remainder of this section thus considers the feasibility of a different approach that focuses on shared empirical implications of the different sets of ethical principles. In particular, one way to understand an AI system is in terms of its operating characteristics: how does the system behave across different environments, including contexts in which it suffers different types of operating failures? A system's operating characteristics ideally provide a relatively complete description of the system's performance and so should be independent of the particular ethical principles advanced by the originators or adopters. Suppose we can determine the operating characteristics that are required, permissible, or forbidden given a particular set of ethical principles. In that case, we can arguably determine ethical interoperability in a straightforward way: take the system and check where it resides in the relevant landscape for the adopter's ethical principles. That is, operating characteristics might provide a way to determine the "extension" of the ethical principles (i.e., the AI systems to which the principles "refer"), and so we could perhaps check for ethical suitability at that level of description. Moreover, operating characteristics are often disclosable to allies or partners, as secrecy concerns are typically about methods and mechanisms, not system behaviors (which are often directly observable in practice).

One might thus hope that ethical interoperability could be solved by the originator disclosing the operating characteristics for an AI system, followed by the adopter simply checking whether those operating characteristics are permissible given the adopter's ethical principles. As noted above, this type of solution is analogous to methods for resolving linguistic disagreements by identifying discrepancies in the extensions of the relevant concepts. This approach is also similar to the strategy of consequentializing different moral theories (Portmore, 2007, 2009) so we can

[2] One might respond that only minimal efforts have been made to develop such a language, and so we should simply work harder. We return to this idea at the end of Sect. 3.

identify discrepancies or disagreements. This latter strategy is useful when we are uncertain whether two ethical theories actually disagree; for instance, when does some deontological view actually disagree with a particular virtue ethic? We can approach this question by consequentializing each—that is, by determining the consequentialist implications of each—to reveal the "behavioral" implications of each theory. Of course, moral theories are more than simply their consequentialist implications. Nevertheless, consequentializing can help reveal operational similarities and differences. Similarly, an AI system is more than simply its operating characteristics, but the common framework of operating characteristics might provide a way to solve the problem of ethical interoperability, even if it is insufficient to say whether two different sets of ethical principles are "really" the same. However, a natural question arises: exactly what are the operating characteristics (of an AI system) that would need to be specified for this approach to work? We consider three possible answers to this question.

### 2.1 Characteristics of the Technology

The most minimal set of operating characteristics would be those that depend on only the AI system itself. Ideally, these would be behaviors or properties that depend only on the zeros and ones of the AI system (plus hardware for robotic systems), such as performance on benchmark datasets or formal verification that the software implements the intended function (Dennis et al., 2016). Of course, defense organizations will rarely, if ever, be willing to share actual source or machine code, even with allies. This approach thus requires agreement on either appropriate performance standards that could be tested by the adopter or a trusted third party, or some other trusted certification that the AI system itself has a particular performance profile. By analogy, an originator's radios can be certified as interoperable with the adopter's systems simply by showing that they can appropriately transmit without generating interference; the adopter does not need access to the radios' inner workings but only the performance characteristics of the radios.

In the context of our running examples, the hope would be that the operating characteristics of the AI systems could be appropriately specified and verified so the adopter (Australia or the US Army) would know whether those AIs satisfy their ethical principles. (We assume that the AI satisfies the originator's ethical principles, else they would presumably not be using it in the first place.) For instance, we might hope that Australia could use specifications of US satellite AI performance on, say, benchmark image datasets to determine whether the US system conforms to their ethical AI principles. Similarly, the US Army could try to use operating characteristics of the IC's AI-plus-sensors (e.g., formal verification of functional properties) to confirm that the system satisfies the DoD ethical use principle of governability (and not just the ODNI principles).

This approach is highly appealing as it would enable the establishment of ethical interoperability using observable or behavioral measures that can be documented in standard formats such as datasheets (Gebru et al., 2021) or model cards (Mitchell et al., 2019). Ideally, ethical interoperability could be established using no

information beyond what the adopter could access simply through regular interaction with the AI system (perhaps with additional trustworthy verifications). That is, this approach could conceivably solve the ethical interoperability challenge largely "for free," without requiring additional effort or disclosure.

Unfortunately, this approach will rarely work precisely because satisfaction of most ethical AI principles requires information beyond purely technological details. For example, governability requires that users can "disengage or deactivate deployed systems that demonstrate unintended behavior" (DoD, 2020), and so satisfaction of that principle depends on the broader sociotechnical structures within which the AI system is deployed. The same AI technology (i.e., the same zeros and ones) could be governable or not, depending on how it is deployed and used, including surrounding systems. In fact, even technologies such as radios are not deemed "interoperable" solely on technological grounds: interoperability of communication technologies also depends on usage practices, language, and many factors beyond just the hardware and software.

### 2.2 Characteristics of the Sociotechnical System

These concerns about a focus solely on the technology suggest that we might be able to establish ethical interoperability on the basis of the AI system plus a description of the relevant sociotechnical system, what we can simply call the "sociotechnical AI system." This richer information set would include the social contexts, use practices, stakeholders, and so on. Secrecy concerns obviously loom somewhat larger if the broader sociotechnical system must be described, but in many cases, the information that would need to be disclosed would be relatively unproblematic. For example, the ability of a user to stop image collection if there are AI image processing system failures (i.e., the ability to Govern the AI system) can be established with quite general and high-level information about the methods and processes used by an agency.

A worry about this approach is the difficulty of providing a precise specification of the sociotechnical system. There is no general-purpose formal language to specify the organizational processes, social norms and practices, contexts of deployment or use, or many other elements of a sociotechnical system. Without such a language, there will be persistent risks of terminological confusion where the originators and adopters have different understandings of particular terms. Those situations can readily lead to the adopters concluding (incorrectly) that the AI system satisfies their ethical principles because they misunderstand the exact structure and dynamics of the sociotechnical system. However, this worry might be surmountable, as the development of structured formal languages and specifications of sociotechnical systems is a current, active research topic (e.g., Chopra & Singh, 2018; Singh, 2014). Although we cannot yet solve the challenge of ethical interoperability using operating characteristics of the sociotechnical AI system, we might hope that this approach will provide a solution in the near future.

Unfortunately, even if we have such a formal specification language, the problem of ethical interoperability will persist. In particular, there are cases in which we will

need information about the lifecycle and history of a sociotechnical AI system in order to determine whether it satisfies the adopter's ethical principles. That is, a description of the sociotechnical AI system at a moment will typically be insufficient to determine whether the system satisfies the usual kinds of ethical AI principles. For example, the DoD principle of governability refers to the "intended function" (DoD, 2020) of the AI system, which can depend on the historical problems and contexts that drove the system's development. The DoD principle of reliability explicitly refers to past events, as it holds that AI systems should have been "subject to testing and assurance…across their entire life-cycles" (DoD, 2020). Similar issues arise with the Australian method's principle of trust, which depends on features of the supply chain or test and evaluation that occurred prior to the attempted cooperation and collaboration. There might be occasional cases in which a description of the sociotechnical AI system at a moment is sufficient to verify that it satisfies some set of ethical AI principles, but those cases are likely to be the exceptions. For the challenge of ethical interoperability, a general solution using operating characteristics will require even more information.

### 2.3 Characteristics of the System's Lifecycle

Given the considerations in the previous subsections, we might consider the opposite extreme: perhaps "operating characteristics" must include information about the entire AI system lifecycle of ideation, design, development, deployment, use, and revision. Of course, such a specification would go far beyond the usual understanding of operating characteristics (hence the scare quotes in the previous sentence). However, the previous examples provide strong evidence that satisfaction of ethical AI principles can only be established with information of this scope and scale. Moreover, proposed operationalizations of ethical AI principles (e.g., from the US DIU) often emphasize the importance of considering decisions at all stages in the lifecycle of an AI system.

This level of information would almost surely be sufficient for the adopter to determine if an AI is acceptable to them, as it would include all of the various choices, constraints, evaluations, and processes used throughout the AI's lifecycle. If one knows all that was done in which contexts to produce the AI (or its outputs), then one should know enough to judge its fit with a set of ethical AI principles. Suppose the Australian military knows all of the testing and evaluation that occurred (plus everything else about its creation and refinement) for a US satellite imagery processing AI. In that case, the RAAF can judge whether the AI output conforms to the Method's ethical AI principles. We thus appear to have a solution to the challenge of ethical interoperability.

Of course, the description of this idea immediately reveals its infeasibility. Almost surely, the originators will be unwilling to share the required level of detail and information. For example, the DIU case studies (Dunnmon et al., 2021) describe various decisions in planning, development, and deployment, but always in high-level, qualitative terms. Secrecy and confidentiality concerns, even between defense and security organizations within the same government, mean that this strategy will almost never be realizable in practice. This problem is arguably exacerbated when we expand from national defense organizations to include the broader defense industry, as those

organizations also have significant intellectual property considerations. There is no reason to think that this extreme kind of information-sharing is feasible, regardless of whether it would provide a solution to the challenge of ethical interoperability.

The core tension throughout this section has been between verifiability and secrecy: adopter verification requires substantial information that falls under the scope of (legitimate) originator secrecy. The originator will, we contend, almost never be willing to share enough information with the adopter for the latter to check if the AI system satisfies the adopter's ethical principles. We suggest that the only practical solution is for the *originator* to determine if the system conforms to the *adopter's* ethical AI principles. One might immediately object that the adopter will never (and should never) accept the originator's claims about satisfaction of the adopter's principles. The originator will typically not share the adopter's interests, ethical perspective, understanding of their principles, or other relevant aspects of the verification process. So there is little reason for the adopter to think that the originator could successfully do this work on their behalf.[3] We agree with this concern but contend that "originator verification" nonetheless can provide a way to make progress on the challenge of ethical interoperability. We turn now to articulating and defending this idea.

## 3 An Alternative Path

Although different sets of ethical principles can offer different judgments about particular AI systems, there is also substantial overlap in terms of which sociotechnical AI systems are ethically permissible. For example, all extant sets of principles agree that it is permissible to use a system for weather forecasting that provides appropriate explanations or transparency, is unbiased in terms of its errors, and so on. Similarly, a satellite-based AI system that solely performed image corrections or enhancements (in ways that did not bias downstream uses of the images) would presumably be permissible under every proposed set of ethical AI principles. Of course, any system could theoretically be used in ways that violate some set of ethical AI principles, which is why an AI's operating characteristics alone are insufficient (as argued in Sect. 2.1). However, there is a wide range of cases that would be deemed ethical under various sets of principles; not every situation involves the challenge of ethical interoperability.[4] That is, we should potentially be able to articulate systems, uses, and contexts that would conform with most or all ethical AI principles from national defense organizations.

In light of these observations, we propose the following framework (and consider objections in the latter part of this section). We assume throughout that each defense organization already has mechanisms for determining whether some sociotechnical AI system satisfies the totality of its own ethical principles. Given such processes,

[3] One might also be concerned that the originator has only limited incentives to do a good job at the verification, regardless of their capabilities. We consider incentives in more detail in the next section.

[4] By analogy, essentially all normative ethical theories would agree that it is permissible to assist someone who is drowning (though they might disagree on whether it is obligatory), even though there are edge cases in which this act might be problematic.

defense organizations who are interested in collaborating should work to identify a set of sociotechnical AI systems that all relevant parties agree are in a zone of "low risk of violating their own set of ethical principles." These will be particular systems, uses, and contexts that each of the collaborating parties agrees are highly unlikely to violate their own particular set of ethical AI principles. That is, the first step would be for defense organizations to determine which potential sociotechnical AI systems would be "ethically safe" in the sense that they would be highly likely to pass their own (internal) ethical evaluations, including their own ways of understanding principles, resolving tradeoffs, and so forth.

We emphasize from the outset that these are *not* necessarily "low risk" AI systems in the usual use of that term. The risk of harm posed by a system is conceptually, and often practically, separable from the risk that the system violates a set of ethical AI principles. In one direction, systems that can cause harm (and so are, in some sense, "risky") can nonetheless conform to one's ethical AI principles. For example, the US Navy's Aegis Combat System,[5] which is used to defend ships against a range of threats, is an AI-powered system that provides decision support for high-consequence, potentially life-or-death actions. The Aegis system is used by Japan, Spain, Norway, the Republic of Korea, and Australia (in addition to the US), and despite being "risky" (in the sense of being capable of significant harm), would likely satisfy any set of ethical AI principles from these nations as it is highly explainable, controllable, transparent, and unbiased. In the other direction, AI systems for lower-consequence domains and decisions could be highly likely to *not* satisfy one's ethical AI principles. For example, a human-in-the-loop AI system designed to support human resources in choosing candidates for promotion could be relatively low-risk in terms of harm but quite biased, opaque, and generally problematic in terms of ethical AI principles. The categories of "low/high risk of violating a set of ethical principles" do not map straightforwardly onto the categories of "low/high risk of harms."

Returning to our initial examples, we suggest that the US Army's use of IC assets does not pose a significant risk of violating the DoD ethical AI principles as long as there is close interaction and coordination (in the broader sociotechnical system) between the two groups. For example, the DoD requirement of governability can be, we suggest, satisfied through monitoring and awareness of the performance of the IC assets, even if the DoD team cannot directly control the AI. (Of course, one would need to confirm that all of the DoD principles are likely satisfied, not just this one.)

Agreements between organizations about the zones of "low ethical principles risk" (Low-EPR) could be established in many different ways. Most obviously, different defense organizations could engage in bilateral or multilateral negotiations about exactly how to characterize the space for those groups. The NATO principles, along with other efforts such as the Partnership for Defense, provide hope that such negotiations could lead to usable, practical agreements. Importantly, these discussions would not require disclosure of any actual methods, algorithms, or systems. Instead, the goal of negotiations would be agreement about the sociotechnical AI

[5] For more information about the Aegis system see: https://sgp.fas.org/crs/weapons/RL33745.pdf

systems that would satisfy all parties' ethical AI principles *if* they were designed, developed, deployed, and used in particular ways.

Moreover, there is no assumption that these discussions would lead to the same Low-EPR zones for all groups of defense organizations. It is possible, perhaps even likely, that the Low-EPR zones would vary depending on exactly which parties are involved, precisely because of differences in their ethical AI principles. For instance, NATO might settle on a different Low-EPR zone than the Five Eyes since different countries, and so different sets of ethical AI principles, would be at the table. These Low-EPR zones could also presumably shift over time as defense organizations better understand the scope and implications of their principles.

Given such agreements, we propose that the originator defense organization can self-certify the sociotechnical AI system as falling into the relevant Low-EPR zone. Further research is needed to determine the best mechanisms by which a self-certification is transmitted to the potential adopter. Regardless, this certification would enable the adopter organization to use the system (or its outputs) with appropriate confidence that it satisfies their ethical AI principles (since everything in the Low-EPR zone is agreed to be highly likely to satisfy each group's ethical AI principles). That is, we have a (partial) solution to the problem of ethical interoperability.[6]

We acknowledge that this framework does not provide a perfect solution. In particular, there can be sociotechnical AI systems that actually would satisfy the adopter's set of ethical principles (and so could be used), but that fall outside of the Low-EPR zone so cannot be collaboratively employed using the self-certification process. However, there is no mechanism that would enable *perfect* ethical interoperability. We contend that it is better to err on the side of caution than to risk having a group use an AI system that violates their ethical principles. If defense organizations are willing to tolerate more risk in this regard, then they can negotiate to expand the relevant Low-EPR zone.

This framework would enable significant progress towards achieving ethical interoperability, but we now consider five potential concerns about its feasibility. First, international negotiations and deliberations about defense AI have been notoriously unsuccessful in recent years, so one might reasonably worry that negotiations about a Low-EPR zone are doomed to failure. There are at least three reasons to hope for success. First, these discussions would be centered on identifying cases that everyone can agree to be low risk for violating ethical principles. In contrast, previous negotiations have focused mainly on cases with potentially high risk of harms (e.g., systems with potentially lethal effects or other weaponized AIs), where there is significantly more disagreement. Second, agreement about a Low-EPR zone of sociotechnical systems would not actually constrain any defense organization in terms of what they create for their own uses. These zones are only to enable ethical collaboration. Third, and most importantly, these negotiations would

[6] There are obvious similarities between this framework and Article 36 weapons reviews, as both involve self-certification that some system is (ethically) permissible for use in international engagements and collaborations. However, there are also numerous *dis*similarities between these frameworks, including different stakeholders (here only the parties to an agreement vs. all nations for Article 36); different incentives (helping allies vs. reassuring adversaries); different standards (low ethical principles risk vs. prohibited by Geneva Conventions); and different mechanisms (part of internal reviews vs. separate review).

occur between parties who presumably want to find a solution to the problem of ethical interoperability (e.g., defense allies, or a military and its suppliers), and so have significant incentives to reach agreement. Agreement on the NATO AI principles provides positive evidence that negotiations around ethical interoperability could succeed.

A second concern is that defense organizations might worry that the act of self-certification could potentially reveal information about capabilities and technologies. Even allies who are trying to cooperate do not want to share absolutely everything (as argued in Sect. 2.3). The originator's declaration that "this system is in the agreed-upon Low-EPR zone (and so the adopter can ethically use it)" conveys additional information, but we suggest that it is relatively minimal information. The originator's offer of AI access has already revealed the existence of a system with various capabilities and the fact that the system conforms to the *originator's* ethical AI principles (else they presumably would not use it themselves). The self-certification declaration thus only reveals the additional information that the system is not in a zone of potential dispute where it might violate the adopter's ethical AI principles. We contend that this is relatively uninformative about capabilities; the substantive information was already conveyed in the offer to collaborate.[7]

Third, an originator organization might declare that a particular AI system is in the Low-EPR zone even when the system is not, or when the originator has not bothered to check. This worry is the specific version of the general concern about any system based on self-certification: what are the incentives for self-certifiers to be accurate and diligent? We agree that failures in self-certification are a potential concern, but we note two reasons to think that honest self-certification is more likely in this situation. First, most (and perhaps all) of the information required for the self-certification would already be obtained through the originator's internal ethical AI certification processes in which they determine that their own ethical AI principles are satisfied by this system (e.g., those required by DoD, 2021). Self-certification is unlikely to require substantial additional effort by the originator; there should not be significant additional costs.[8] Second, these self-certifications would occur between allies, so there would be significant potential risks if the originator engaged in deceptive behavior. While defense organizations could still engage in deception, such actions would risk harm to presumably valuable relationships. This motivation is particularly salient for defense contractors and suppliers who want national defense organizations to adopt their AI systems.

A fourth concern is that some defense organizations have proposed ethical AI principles that emphasize national control (e.g., Australia, France), and we earlier argued

[7] In theory, additional information could be conveyed if the originator declared the AI system to be in the Low-EPR zone for adopter *X* but not for adopter *Y*, as one could look to see what kinds of systems reside in the former Low-EPR zone but not the latter. However, this inference assumes that one knows the two declarations and the two Low-EPR zones, but our framework does not require these to be made public. In our view, ethical interoperability involves only the originator and the adopter, not the rest of the international community.

[8] In fact, we think that it is plausible that agreements about Low-EPR zones may lead developers to sometimes build with that self-certification in mind, rather than only thinking about their own group's ethical AI principles. Of course, that possibility is speculative, but it does suggest that this framework could lead to more ethical defense AI, not only more ethical interoperability.

that this commitment implies that these organizations could (almost) never be adopters of another group's sociotechnical AI system (or its output). If the adopter must have sovereign control in some strong sense, then the only AI systems in their Low-EPR zone would be those that the originator cedes to them (and there are likely to be vanishingly few of those). We concede that principles of sovereign control might preclude participation in this framework, though such principles arguably preclude *any* cooperation involving AI. That is, the issue is not with our proposed framework but with the very idea of cooperation when one party insists on sovereign control. More importantly, though, our framework is arguably consistent with the *justifications* that underlie principles of national control. Those justifications primarily center on the importance of building capabilities and reducing dependence; that is, ethical AI should preserve or enhance the nation's independence in defense AI. Those goals could readily be incorporated into agreements about Low-EPR zones. For example, an adopter nation that is concerned about reducing (problematic) dependence could require that a sociotechnical AI system is Low-EPR only if there are clear mechanisms to ensure continued access. Or an adopter that is concerned about increasing capabilities could require Low-EPR zone AI to be transparent with regards to methods. Of course, other nations might reject these requirements, thereby shrinking the Low-EPR zone. In the worst case, a nation might find itself unable to reach any agreements about Low-EPR zones. However, the worst case would occur exactly when that nation is unable to ensure that others' AI systems are likely to satisfy their ethical principles, and so they ought not use those systems in the first place. That is, this framework would actually enable a nation to continue living up to its ethical principles.

Fifth, one might object that this framework would enable each defense organization to continue using its own set of ethical AI principles, rather than those groups being forced to negotiate towards a shared standard. One might hope that we could develop shared understandings of precisely what defense AI is ethically permissible for what uses in which situations. If we develop and implement a framework for ethical interoperability, then we remove some of the pressure to reach that ideal, as we would have a mechanism for defense organizations to resolve their disagreements without developing shared language or principles. However, this ideal may be unreachable: even the NATO AI Principles, while a step in the right direction for creating a common baseline, acknowledge that national obligations and commitments exist and must be respected. More practically, calls to develop "shared terminology," not mere lists, have sometimes been used as a delaying tactic in negotiations. In other domains (e.g., cybersecurity), progress has come primarily when parties have focused on agreements about specific cases rather than abstract principles. We contend that the same is true here and that our proposed framework would likely increase the ethical use of defense AI rather than being an impediment.

## 4 The Scale of Ethical Interoperability

We have focused on AI in defense and security contexts, as those involve both explicit statements of ethical principles by relevant parties and frequent need for collaboration (at least, between allies and friends). These contexts make vivid the need

and difficulty of ethical interoperability and so are particularly useful in explicating the challenge and possible solutions. However, the challenge of ethical interoperability is clearly not restricted to defense and security contexts, but can readily arise whenever one party wants to use AI technology or outputs that originate with another party. Most obviously, advanced AI systems are increasingly being developed by a small group of companies who then provide "AI as a service" to other companies, nonprofits, or governments. The adopters or recipients of those services will often have different ethical principles than the originators or developers, and therefore must confront the challenge of ethical interoperability.

Consider some non-defense examples.[9] A company wants to use the Microsoft Azure AI platform to generate recommendations for their customers. Microsoft has a set of responsible AI principles and practices,[10] and the methods they release onto Azure presumably satisfy Microsoft's principles. The relevant question, though, is whether those methods satisfy the ethical principles of the company *using* Azure, not Microsoft itself. Alternatively, consider a social media platform that wants to use the Google/Jigsaw Perspective API system[11] that detects toxicity in messages or other content. Google has ethical AI principles[12] that presumably guide their work, but the relevant question is whether Perspective API satisfies the *platform's* ethical principles. Those same Google principles presumably also cover the diagnostic AI systems developed by Google Health (e.g., DermAssist[13]), but the challenge of ethical interoperability is whether that diagnostic AI conforms to the ethical guidelines and principles of local healthcare providers who want to use it. Alternatively, consider a company that aims for more ethical hiring practices and decides to partner with pymetrics, which provides additional measures and methods to support hiring diverse, qualified candidates. pymetrics has undergone a third-party audit (Wilson et al., 2021) to confirm that their system follows their own ethical principles,[14] but the relevant question here is whether their system satisfies the hiring company's ethical principles. As a final example, the UN Food and Agriculture Organization (FAO), which endorses the principles in the 2020 Rome Call for AI Ethics,[15] aims to use AI systems from IBM and Microsoft to improve agricultural yields across the world (FAO, 2020). The challenge of ethical interoperability here is whether and how FAO can have appropriate guarantees from Microsoft and IBM that their AI systems also conform to the Rome principles, not only those of the companies themselves.

[9] These examples name particular companies purely for specificity, not because of any concerns about the ethical nature of their technologies or uses. In fact, we believe that these companies should all be lauded for their commitments to ethical and responsible AI through both principles and practices.

[10] https://www.microsoft.com/en-us/ai/responsible-ai

[11] https://www.perspectiveapi.com

[12] https://ai.google/principles/

[13] https://health.google/consumers/dermassist/

[14] https://www.pymetrics.ai/mission

[15] https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/CALL%2028%20febbraio/AI%20Rome%20Call%20x%20firma_DEF_DEF_con%20firme_.pdf

These examples suggest that ethical interoperability is a pervasive problem across almost all sectors. The drive to develop sets of ethical AI principles has provided clear benefits in terms of focusing attention on the ways in which technology can be misused or poorly used. However, that same drive has also created a considerable, relatively unnoticed challenge to (ethical) cooperation and collaboration.

These examples provide further support for the negative arguments of Sect. 2 that operating characteristics cannot provide a universal language for ethical interoperability. For example, a specification of the Perspective API algorithm(s) to flag content as 'toxic' provides insufficient information to know whether the system is "governable" by the adopting social media platform, as that specification underdetermines the level of control that is available. Proposals to include reliability or other requirements in acquisition and purchase contracts may be sufficient for legal purposes, but almost never for ethical purposes. At the other extreme, pymetrics would clearly be unwilling to release all relevant details about their system's history, testing, and development, as they did not even provide that information to the third-party audit team (Wilson et al., 2021). The core tension between privacy (of the originator's AI) and verifiability (of satisfaction of the adopter's ethical AI principles) is not defense-specific. The reasons for the tension might differ in other sectors (e.g., originator's desire for privacy to protect their intellectual property, rather than national security), but regardless, appeal to operating characteristics will again not work.

The positive proposal in Sect. 3 is suggestive, but does not perfectly translate to non-defense settings. The general idea of originator self-certification that the AI system falls into an agreed Low-EPR zone could still work in the abstract, but the negotiations to establish those zones would have a radically different form in non-defense settings. There are many more actors in non-defense settings, and there are much greater power imbalances between them. For example, far more companies want to use Microsoft Azure AI than national defense organizations who want to collaborate with the US DoD. Moreover, the adopting companies have relatively little bargaining power when they engage with Microsoft. It would be impractical to expect every company *C* to engage in bilateral negotiations with Microsoft to establish a Low-EPR zone solely for *C's* use of Azure, but it would also be inappropriate for Microsoft to provide no guidance.

For these reasons, we propose that professional societies and trade organizations should play a crucial role in addressing ethical interoperability in non-defense settings. That is, we propose that groups such as IEEE, Partnership on AI (PAI), or the National Association of Manufacturers should engage in deliberative processes to develop Low-EPR zones for their sectors and members. These zones would presumably be non-binding, as these groups have no particular enforcement powers. Instead, they would function like industry standards where adherence would enable companies to be confident that they were adhering to their own ethical principles, even when using AI systems developed by others.[16] These zones would also

[16] Industry standards sometimes also serve a legal function, as adherence to industry standard practices can provide a presumptive defense against negligence claims. We do not expect that Low-EPR zones would currently serve such a function, as there are (to our knowledge) no sets of legally binding ethical AI principles at the moment.

potentially support contractual language to establish penalties if a company's self-certification were found to be fraudulent. Ideally, the existence of such "standard" Low-EPR zones would incentivize companies to develop AI technologies that reside in them, potentially leading to more ethical AI systems in general.

## 5 Conclusion and Future Research

The challenge of ethical interoperability is particularly salient for defense organizations, but as ethical AI principles are increasingly adopted, this challenge will arise for essentially any group that wants to use AI systems or outputs from another group. Regardless of sector, developers and users of AI systems should embrace the challenge of ethical interoperability, as the alternative—adopters do not worry about whether the originator adhered to the former's ethical AI principles—would open the door for problematic "ethical outsourcing." Adopters should not be able to use systems that violate their own ethical AI principles simply by getting some other group to develop the system for them. If organizations take seriously their ethical commitments, they must ensure that every AI they use, regardless of source, conforms to their principles.

At the same time, some open research and application questions remain about ethical interoperability, including our positive proposal. First, we have argued that ethical interoperability will require some self-attestation component, but auditing and verification of performance standards can also play a role. One challenge is to understand better the advantages and shortcomings of different balances of self-attestation and verification: in particular, how much verification can be used in real-world cases, thereby reducing the scope of the self-attestation? Relatedly, ongoing operationalizations of high-level defense AI principles may impact the acceptability and value of self-attestations, depending on whether different organizations operationalize in similar or different ways. Second, we have provided some simple examples of AI systems and uses that are highly unlikely to violate an organization's set of ethical AI principles, but it would be helpful to have a more general characterization of the space of (likely) Low-EPR cases. Third, although we earlier argued that EPR and "regular" risk are distinct, they are related in some cases, and so there are potential connections between our positive proposal and various emerging risk-based legal and regulatory frameworks (e.g., EU AI Act or US NIST Risk Management Framework, both still in draft stage at the time of this writing). And fourth, we recognize that there may be cases in which an organization needs to use an AI system from another actor, even when it does not fall into a negotiated Low-EPR zone. In these cases, ethical interoperability again poses a challenge, and so methods or frameworks for such situations will be needed.

Ethical interoperability poses a significant challenge, particularly in the defense and security sector, but we need not resign ourselves to "anything goes" nor tolerate ethical outsourcing. The positive proposal developed here offers a path towards such assurances, thereby providing a potential pathway to enable organizations to operationalize their ethical AI principles and meet the challenge of ethical interoperability.

**Acknowledgements** Thanks to Thilo Hagendorff for helpful conversations about this challenge. Two reviewers provided valuable comments and critiques of earlier versions of this paper.

**Funding** DT was partially supported by the Swiss Drone and Robotics Centre.

## Declarations

**Conflict of Interest** DD is a member of the Editorial Board for Digital Society.

**Disclaimer** The views in this paper are solely those of the authors and do not represent the position of any other group or agency.



## References

AI Task Force (AITF). (2019). Artificial intelligence in support of defence. Paris, France: French Ministry of Armed Forces.

Chopra, A. K., & Singh, M. P. (2018). Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp 48–53).

Defense Innovation Board (DIB). (2019). AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. Accessed Date: 18 Feb 2022

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1–14.

Department of Defense (DoD). (2020). DOD adopts ethical principles for artificial intelligence. https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/. Accessed Date: 18 Feb 2022

Department of Defense (DoD). (2021). Memo outlines DOD plans for responsible artificial intelligence. https://www.defense.gov/News/News-Stories/Article/Article/2640609/memo-outlines-dod-plans-for-ethical-artificial-intelligence/. Accessed Date: 18 Feb 2022

Devitt, S. K., & Copeland, D. (2021). Australia's approach to AI governance in security and defence. arXiv: 2112.01252. Accessed Date: 18 Feb 2022

Devitt, K., Gan, M., Scholz, J., & Bolia, R. (2020). A method for ethical AI in defence. Canberra, Australia: Australian Department of Defence. https://www.dst.defence.gov.au/sites/default/files/publications/documents/A%20Method%20for%20Ethical%20AI%20in%20Defence.pdf. Accessed Date: 18 Feb 2022

Dunnmon, J., Goodman, B., Kirechu, P., Smith, C., & van Deusen, A. (2021). Responsible AI guidelines in practice: Lessons learned from the DIU portfolio. https://www.diu.mil/responsible-ai-guidelines. Accessed Date: 25 Mar 2022

Food and Agricultural Organization (FAO). (2020). Artificial Intelligence best-practices in agriculture can help bridge the digital divide while tackling food insecurity. Sept. 24, 2020. https://www.fao.org/news/story/pt/item/1309630/icode/. Accessed Date: 24 Feb 2022

Feickert, A. (2021). The Army's Project Convergence. https://crsreports.congress.gov/prodauct/pdf/IF/IF11654. Accessed Date: 18 Feb 2022

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020–1).

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp 5–17). Springer, Cham.

Freedberg S. J., Jr. (2020). 'Improvised mode': The Army network evolves in Project Convergence. https://breakingdefense.com/2020/09/improvised-mode-the-army-network-evolves-in-project-convergence/. Accessed Date: 24 Feb 2022

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Hoehn, J. R. (2022a). Advanced Battle Management System (ABMS). https://sgp.fas.org/crs/weapons/IF11866.pdf. Accessed Date: 18 Feb 2022

Hoehn, J. R. (2022b). Joint All-Domain Command and Control (JADC2). https://sgp.fas.org/crs/natsec/IF11493.pdf. Accessed Date: 18 Feb 2022

Ministry of Defence (MoD). (2022a). Ambitious, safe, responsible: Our approach to the delivery of AI-enabled capability in Defence. Policy paper. June 15, 2022. https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/. Accessed Date: 6 Jul 2022

Ministry of Defence (MoD). (2022b). Defence Artificial Intelligence strategy. *Policy Paper*. June 15, 2022. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy. Accessed Date: 6 Jul 2022

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp 220–229).

North Atlantic Treaty Organization (NATO). (2021). Summary of the NATO Artificial Intelligence Strategy. Retrieved from https://www.nato.int/cps/en/natohq/official_texts_187617.html. Accessed Date: 25 May 2022

Office of the Director of National Intelligence (ODNI). (2020a). Artificial intelligence ethics framework for the Intelligence Community. https://www.intelligence.gov/images/AI/AI_Ethics_Framework_for_the_Intelligence_Community_1.0.pdf. Accessed Date: 18 Feb 2022

Office of the Director of National Intelligence (ODNI). (2020b). Principles of artificial intelligence ethics for the Intelligence Community. https://www.intelligence.gov/images/AI/Principles_of_AI_Ethics_for_the_Intelligence_Community.pdf. Accessed Date: 18 Feb 2022

Portmore, D. W. (2007). Consequentializing moral theories. *Pacific Philosophical Quarterly, 88*(1), 39–73.

Portmore, D. W. (2009). Consequentializing. *Philosophy. Compass, 4*(2), 329–347.

Reeves, T. (2021). Military interoperability in the intelligent age of warfare. Deloitte Center for Government Insights report. https://www2.deloitte.com/global/en/pages/public-sector/articles/military-interoperability-in-the-intelligent-age-of-warfare.html. Accessed Date: 25 May 2022

Royal Australian Air Force (RAAF). (2019). At the edge fifth generation Air Force: Our human edge in the information age. https://view.publitas.com/jericho/at-the-edge/page/1. Accessed Date: 18 Feb 2022

Singh, M. P. (2014). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(1), 1–23.

Stanley-Lockman, Z. (2021). Responsible and ethical military AI: Allies and allied perspectives. https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/. Accessed Date: 18 Feb 2022

Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. *Philosophy & Technology, 34*, 1707–1729.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp 195–200).

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp 666–677).