J Multidisciplinary Scientific Journal

MDPI

Article

The Ethical Assessment of Autonomous Systems in Practice

Daniel Trusilo [1,*] and Thomas Burri [2]

[1] School of Economics and Political Science, University of St. Gallen, 9000 St. Gallen, Switzerland
[2] Law School, University of St. Gallen, 9000 St. Gallen, Switzerland; thomas.burri@unisg.ch
* Correspondence: daniel.trusilo@student.unisg.ch

Abstract: This paper presents the findings of a study that used applied ethics to evaluate autonomous robotic systems practically. Using a theoretical tool developed by a team of researchers in 2017, which one of the authors contributed to, we conducted a study of four existing autonomous robotic systems in July 2020. The methods used to carry out the study and the results are highlighted by examining the specific example of ANYmal, an autonomous robotic system that is one component of the CERBERUS team that won first place in DARPA's Subterranean Challenge Systems Competition in September 2021.

Keywords: robotics; artificial intelligence; autonomous weapons systems; international humanitarian law; applied ethics

check for updates

Citation: Trusilo, D.; Burri, T. The Ethical Assessment of Autonomous Systems in Practice. *J* **2021**, *4*, 749–763. https://doi.org/10.3390/j4040051

Academic Editors: Ugo Pagallo and Massimo Durante

Received: 7 October 2021
Accepted: 1 November 2021
Published: 10 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In perhaps one of the most profound moments in the debate about the development and use of autonomous robotic systems and their ability to be used as weapons, the UN Panel of Experts on Libya published in June 2021 the first officially recognized instance of a robotic system using artificial intelligence (AI) to target humans autonomously [1]. The development of such technology, known as autonomous weapons systems (AWSs), presents a range of problems related to fundamental notions that international humanitarian law (IHL) is built on [2], respect for human life [3] (p. 110), and the role of humans in the development and operationalization of autonomous robotic systems.

However, autonomous robotic systems also have enormous potential to do good. A system optimized to operate autonomously could conceivably enhance traditional human search and rescue efforts. In disaster relief operations, an autonomous system could help locate trapped victims in a collapsed structure, carry out damage assessments in inaccessible terrain, or assist in a Fukushima-type nuclear disaster while keeping humans out of harm's way [4].

Whether an autonomous robotic system is designed as a weapon or to support humanitarian operations, there is potential to cause harm to humans. However, the law on autonomous systems remains unsettled and vague at present. Due to the lack of legal clarity, autonomous robotic systems are, at the present moment, best assessed from an ethical perspective.

This paper aims to explain how ethical assessments of autonomous robotic systems are best conducted in practice. The paper draws on an ethical assessment tool we previously developed [5]. In the summer of 2020, we applied this tool to assess a series of autonomous robotic systems. A visual representation of the full results of our ethical evaluations can be observed in Appendix A. The paper conveys the essence of our experience from this practical application.

The paper is structured as follows. Section 2 introduces ethical assessments of autonomous systems designed for security purposes, i.e., for deployment in armed conflict and for disaster relief, which is our focus. By discussing the legal framework and two other ethical assessment tools, this section illuminates the background of our research. Section 3 presents the ethical assessment tool that one of us created in collaboration with a group of researchers from several disciplines (henceforth the "tool" or the "Schema") [5] and some of its limitations. Section 4 provides an account of the series of assessments we performed with this tool in 2020. This section does not comprehensively present or discuss all the data we gathered but rather some flashpoints and discoveries we made in the ethical assessments. It illustrates them by using a quad-pedal autonomous robot called ANYmal, which we assessed. Section 5 highlights lessons from our research to improve our tool and ethical assessments more broadly. Lastly, Section 6 offers a brief conclusion.

Our overarching research goal is to address the need to pragmatically inform the research and design, procurement, and operationalization stages of autonomous system development in the absence of clear laws or regulations. We offer a tool tested in practice to systematically identify ethically and legally high-risk properties of autonomous robotic systems to reach this goal. However, rather than providing straightforward answers to ethical questions, our research aims at fostering discussion about risks that a system poses and how a system can be retooled to mitigate potentially harmful properties. The act of discussing these risks may be one of the most critical steps in creating technology for the benefit of humanity.

2. Background

This section addresses the law applicable to autonomous robotic systems designed for security purposes (Section 2.1) as ethical assessments must respect the legal framework. It discusses IHL, for it governs AWSs and peacetime law, which applies to disaster relief and law enforcement systems. The section then features two other ethical assessment tools (Section 2.2) to provide context for our tool, discussed in Section 3.

2.1. Law Applicable to Autonomous Security Systems

At the international level, the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) established a Group of Governmental Experts in 2016 to identify and define the characteristics of AWSs. An AWS can be understood according to the US Department of Defense's definition as a "weapon system that, once activated, can select and engage targets without further intervention by a human operator" [6] (p. 13). The Group of Governmental Experts affirmed 11 guiding principles in its 2019 meeting [7]. These guiding principles are a positive step in establishing international norms. However, "humanity is at the threshold of a new technology that could fundamentally change our relationship with war" [8] (p. 359), and little progress has been made in the CCW discussions as to how such guiding principles can be operationalized.

Thirty countries have called for an outright ban on AWSs [9]. Many activists and researchers have labeled their development a dangerous and potentially destabilizing third revolution in warfare after gunpowder and nuclear arms [10]. Advocates for the prohibition of autonomous systems as weapons of war refer to the preemptive ban on blinding lasers as a precedent to call for a similar ban on the development of AWSs, arguing that prohibitions are feasible without negatively impacting the peaceful uses of emerging technologies [11]. The Campaign to Stop Killer Robots and ethicists such as Robert Sparrow argue from a deontological perspective that allowing life or death decisions to be made by machines crosses a fundamental moral line and would "profoundly disrespect the humanity" of those targeted [3,12].

One of the arguments supporting a ban relies on concerns raised by the Martens Clause, codified in Article 1(2) of Protocol 1 to the Geneva Conventions. The Martens Clause refers to the importance of the principles of humanity and the dictates of public conscience in the absence of existing law relating to a new weapon [13,14]. Proponents of a ban state that autonomous weapons systems can cause harm in inhumane or socially unacceptable ways and should, therefore, be forbidden [15].

A further argument in favor of a ban states that since IHL is intended to address human behavior in war, systems capable of selecting and engaging targets independently of a human operator make IHL difficult, if not impossible, to apply. Under the current legal framework, the challenge of holding a human liable for the actions of an AWS is said to create an accountability gap [16], which Santoni de Sio and Mecacci explain is just one form of four responsibility gaps presented by intelligent systems [17]. Proponents of a ban on AWSs highlight that the concept of meaningful human control is not only central to the debate about AWS but that the inability to ensure human control of previous weapons technologies has motivated past disarmament treaties [9]. Specifically, treaties banning the use of victim-activated landmines, biological weapons, and chemical weapons serve as precedents for bans on AWSs.

Other parties oppose categorical bans that may impact the rapidly developing ecosystem of autonomous systems. For example, the US Department of Defense (DOD) has an interest in seeing autonomous systems developed for both humanitarian and combat operations and multiple Defense Advanced Research Projects Agency (DARPA) public challenges have focused on disaster relief. However, DARPA is a defense department agency, and the primary mission of the DOD is to provide the military forces needed to deter war and ensure national security [18]. The position opposing categorical bans is often supported with utilitarian arguments in which the consequences are the determining factor. The utilitarian argument for AWSs is exemplified by researchers such as Ronald C. Arkin, who argue that it is possible to develop an architecture that governs ethical behavior by mathematically encoding a system's perceptions and behavioral responses [19]. This perspective implies that perfect ethical behavior can be programmed; therefore, one could entrust an autonomous system to be deployed in specific environments without the potential to cause undue harm. Such a theory supports the argument that the use of autonomous systems will result in the most morally acceptable outcome. However, ethical norms are complex, change over time, and are contextual; therefore, they would be challenging to code [5], rendering a utilitarian argument only valid under precise conditions. It is also necessary to note the amoral opposition to banning AWSs from the commercial sector, which does not publicly participate in the debate surrounding AWSs but is interested in pursuing government contracts [20].

Existing IHL applies to AWSs as to any other weapon used in armed conflict. IHL notably includes the principles of distinction, proportionality, and humanity, which potentially curb some of the excesses conceivable with AWSs. To implement and apply these principles, Article 36 of Additional Protocol I to the 1949 Geneva Conventions [21] requires states to conduct a domestic weapon review before acquiring or adopting a new weapon, means, or method of warfare. In the concluding chapter to the 2016 book *Autonomous Weapon Systems: Law, Ethics, Policy*, Nehal Bhuta, Susanne Beck, and Robin Geiss clarify that "a weapons review . . . is primarily focused on the question whether a new weapon could be used in conformity with LOAC (the law of armed conflict) as it currently stands," and should be seen as complementary to international legal and ethical limits to autonomy especially given the complex legal and ethical questions posed by AWSs [22]. An Article 36 weapons review is itself is not without challenges [23]. Moreover, such reviews focus on the use of force. However, robots multiply capacities other than the use of force. They tend to support, advise, and extend humans more broadly. Therefore, existing weapons review processes may need to be improved to account for the challenges presented by AWS [24].

For example, intelligent collective behavior or swarm strategies in which multiple systems are linked offer novel capabilities and the possibility of emergent properties [25]. This strategy of networking autonomous systems as systems of systems is one of the most promising fields of AI R&D and is being incorporated in defense platforms currently under development, such as the Airbus Future Combat Air System [26]. In Will Knight's May 2021 *Wired* article "The Pentagon Inches Toward Letting AI Control Weapons," when discussing swarm technology in warfare, General John Murray, who leads U.S. Army Futures Command, is quoted as asking an audience at the U.S. Military Academy, "Is it within a human's ability to pick out which ones have to be engaged and then make 100 individual decisions? Is it even necessary to have a human in the loop?" [27]. Existing weapons review processes are not well suited to addressing such questions. Therefore, new, practical approaches are needed.

Swarm technology is also disruptive to the regulatory paradigm and processes through which governance structures are created [28]. A practical, multidisciplinary approach involving lawyers, engineers, computer scientists, and operators is critical to address the challenges and concerns presented by such systems [29]. One benefit of our tool, which we will describe in more detail in Section 3, is that it requires critical interdisciplinary discussions.

For admission to peacetime commercial markets, systems must pass tests other than the weapons review under Article 36. The purpose of such tests is to verify compliance with diverse norms applicable under national laws, including technical norms, e.g., product safety requirements and standards [30], and non-technical norms, such as fundamental and human rights. However, the law and the tests that need to be passed evolve as technology advances. Traffic laws, for instance, are amended to enable highly automated driving [31]. Another example of evolving law is the EU's recently proposed legislation to regulate AI [32–35]. As AI is becoming a more prominent element in robotics with high-level autonomy, including robots designed for disaster relief [36], producers of such robots need to consider compliance with the new EU regulation, particularly the regulation's multi-layered obligation to manage significant risks generated by AI.

Concerning autonomous systems in the security domain, both IHL and peacetime laws are generally relevant. While peacetime law itself sometimes states that it does not apply to weapons systems [32], general international law requires compliance with peacetime law when weapons are deployed in armed conflict, even if in rather vague terms [37,38]. Compliance with both IHL and peacetime law makes sense also because modern robotic platforms may be used variably in armed conflict or law enforcement and disaster relief. Since particular arms, such as pistol caliber firearms, can also be used outside of armed conflict, Article 7 of the recent Arms Trade Treaty obliges states to assess the risk that arms could be used to commit violations not just of IHL but also of international human rights law [9,39,40].

2.2. Related Ethical Assessment Tools

Other applied ethics approaches offer well-developed examples of how to evaluate and inform complicated topics in which legal norms have not yet crystallized. For instance, principlism relies on the notions of autonomy, beneficence, non-maleficence, and justice to evaluate bioethical decisions related to topics such as genetic modification, euthanasia, and medical testing [41]. There are two ethical assessment tools other than the Schema we applied in our research that are particularly relevant. On the one hand, Aimee Van Wynsberghe developed the care centered framework and care centered value sensitive design (CCVSD) methodology [42]. Designed to apply to healthcare robots and then expanded to include a broader range of service robots [43], Van Wynsberghe's work is an example of a methodology and related framework utilizing the applied ethics practices of value sensitive design (VSD) and care ethics. Van Wynsberghe addresses Peter Asaro's call for a comprehensive approach to robot ethics [44], applying care ethics to normatively ground the values that can be used to inform the design and operationalization of healthcare and other types of service robots.

One could imagine the methodology and framework developed by Van Wynsberghe applicable to search and rescue robots as they are a category of service robots; however, Van Wynsberghe's work focuses on the reciprocal relationships of care [43]. Therefore, it is not well-suited to address the capabilities of search and rescue systems, nor is the CCSVD methodology applicable to AWSs at all as such systems are, by definition, designed to cause harm. However, Van Wynsberghe's research reinforces the urgency of the proposed project as she makes the compelling argument that "there are no (universal) guidelines or standards for the design of robots outside the factory" [42] (p. 412).

On the other hand, the value sensitive humanitarian innovation (VSHI) guidance framework developed by Ning Wang et al. is another related set of interdisciplinary tools currently being advanced to provide applied ethics guidance for drone technology in humanitarian operations [45]. Such an approach can aid the decision-making process related to the deployment of unmanned aerial vehicle (UAV) technology, addressing the gap between existing principle-oriented recommendations and operational considerations. The VSHI framework aims to involve many stakeholders to apply ethics pragmatically and inform the development and use of new technology. However, the VSHI framework is focused on UAVs for use in humanitarian operations; it does not specifically aim to address autonomous capabilities or the associated ethical issues. Also, it does not apply to the development or use of AWSs. However, it does offer insights into making such a tool thorough and practical for a broad range of stakeholders.

Both the CCVSD-approach and the VSHI-approach serve to increase the trustworthiness of the technology to which they apply. In the case of AWSs, the notion of meaningful human control, which is debated within the CCW, is similarly anchored in trust. When systems are designed to operate at faster than human speeds, trust in them is essential, especially in the absence of clear regulations. In a 2018 article, Heather Roff and David Danks argue that "trust" is a critical factor in the successful acquisition, training, and deployment of any weapon system, be it a conventional weapon, such as a rifle, or a near-future autonomous system capable of contextual learning and decision making [46]. Roff and Danks make the case that trust in the case of an autonomous system refers to a system operator's understanding of the system's values, beliefs, and dispositions. This trust is required for the ethical deployment of autonomous systems so that a person employing such a system can rely on that system without being able to predict its behavior in the complex reality of conflict environments.

In our view, trust is critical in all sensitive operations involving autonomous robotic systems, including search and rescue operations in humanitarian emergencies. Therefore, our research uses applied ethics to increase trust in the systems assessed. Our tool identifies properties of specific systems that are ethically or legally problematic with respect to informing discussions of risks to commonly held values and beliefs.

However, in contrast to the two tools discussed above, our tool focuses on ethical and legal issues of systems that present autonomous capabilities at least to a minimal degree. It is also designed to apply to security systems, i.e., disaster relief systems and weapons systems. However, nothing inherent in its design limits its use to systems designed for these use cases. In 2019 Anna Jobin, Marcello Ienca, and Effy Vayena conducted an analysis, which found 84 documents containing principles and guidelines on ethical AI, including their application in robotics and autonomous systems [47]. However, to our knowledge, our evaluation tool is the only framework designed to go beyond principles and guidelines and practically identify ethically and legally problematic properties of autonomous robotic systems designed for security applications. The specifics of our tool shall be discussed in detail in the next section.

3. Our Approach to Evaluating Specific Aspects of Existing Systems

This section introduces the ethical assessment tool that one of us created together with a group of researchers in 2017 (the Schema) and which we used to conduct assessments of four systems in July 2020 (the assessment of the four systems is discussed in Section 4). This section also briefly explains some of the limitations of our approach.

The Schema was intentionally designed to be relevant to security systems, including disaster relief, law enforcement, and weapons systems, but there is nothing inherent in its design that restricts its use to only the security domain. The first step when applying the Schema verifies that a system is indeed a robotic system. To make this determination objectively, our framework asks whether a system has, to a minimal degree, the following factors: (1) sensing, (2) computing, (3) effecting, and (4) communication capacities.

Next, to determine if a robotic system is autonomous to a degree which warrants the application of the complete framework, a composite definition of autonomy is applied by rating a system's following factors: (1) level of autarchy, (2) independence from human control, (3) ability to interact with the environment, (4) capacity to learn, and (5) level of mobility. Each of these five dimensions of autonomy is evaluated on low, medium, or high scales. A system determined to be "low" in all five dimensions listed above would not fall into the framework's scope. If, however, a single dimension is rated "high" for a particular system, that system is considered autonomous to the degree that warrants the complete evaluation. Systems that are not clearly autonomous according to these dimensions but are also not clearly non-autonomous in one or more dimensions must be considered on a case-by-case basis.

Once a system has been determined to be an autonomous robotic system, the framework addresses whether or not the system is designed with capabilities for which its primary, intended function is to cause harm. If a system is not designed to cause harm intentionally, twenty-nine aspects are analyzed. If, on the other hand, it is determined that a system is designed with a capability for which its primary function is to cause harm, an additional eight criteria are triggered, resulting in an analysis of thirty-seven total aspects.

When evaluating the 29 (or 37) criteria, each criterion is informed by core questions. For example, for the criteria of "physical safeguards," the core questions are as follows: "To what extent do physical safeguards exist that ensure the operator(s) of the system or persons exposed to the robot cannot interfere with mechanical parts of the robots (e.g., rotor protection)? Alternatively, if they can, do safeguards provide sufficient warning from potential dangers?" [5]. Next, for each criterion, an explanation is provided to assist in the determination of a weight and rating. The weight factor is designed to evaluate the relevance of the criteria for the system being evaluated and can be designated as low, medium, or high. In the case of the physical safeguards criterion described above, the weight is determined as follows: "The more moving parts a robot has, and the more kinetic energy the movements of those parts involve, the more relevant is this criterion." Next, a rating of green, amber, red, or grey is assigned according to a detailed series of criteria-specific descriptors. In the case of the example above, a system would be given a "red" rating if "(t)here are clear risks that the robot can harm the operator(s) due to the absence of physical safeguards." The resulting weight and rating values can then be analyzed to determine the risk level that the system assessed poses under each criterion.

Utilizing simplifications and idealizations to order the situationally specific configurations of individual systems into analytical models, as described above, allows us to evaluate complex systems pragmatically. Such research produces case-specific explanations creating what Patrick Jackson labels "analytical narratives" that explain in a practical sense what exists [48]. These analytical narratives allow us to highlight the ethical and legal risk levels of particular properties.

However, the use of analytical narratives results in certain limitations. It means that our findings are limited to concrete examples in time. Using this method, we are not providing generalizable rules applicable to all autonomous systems. Additionally, due to our method, the evaluation is limited to the pre-defined list of criteria in our tool, which may not account for all sources of legal and ethical risk. To mitigate this limitation, we used semi-structured interviews with individual system subject matter experts. The interviews are guided by the core questions and descriptions particular to each criterion as detailed in the Schema. This method allows us to collect as much information as possible about each system in a non-restrictive manner. There is the possibility that an aspect is omitted that would result in the identification of a potentially unfamiliar yet problematic system property; perhaps interviewees do not think to discuss a particular property as they are unfamiliar with what may be legally or ethically concerning. However, our method does not aim to provide an exhaustive checklist of potentially problematic aspects. Instead, its merits lie in the open-ended discussion about ethical and legal concerns with roboticists, engineers, and others involved in the research and design of autonomous robotic systems. Such interdisciplinary discussions raise awareness and encourage interviewees to consider ethical aspects thoroughly.

4. Assessment in Practice

We were able to study four autonomous systems as they were tested as part of the Advanced Robotic Capabilities for Hazardous Environment (ARCHE) exercise in July 2020.

The training facility used for the ARCHE exercise was an internationally standardized search and rescue training facility that replicates various environments (collapsed structures, fires, floods, and uneven terrain) used for training and certifying human search and rescue teams.

The four systems we evaluated were as follows: (1) a four-legged system designed to compete in the US Defense Advanced Research Projects Agency (DARPA) subterranean challenge as part of a multi-modal autonomous search operation, named ANYmal; (2) a fixed-wing unmanned aerial vehicle (UAV) designed to autonomously target and intercept other UAVs using a kinetic "net-gun," named Mobula; (3) a 13-metric ton excavator designed to dig trenches and construct walls autonomously, named Armano; and (4) a quadcopter designed to fly and search for radioactive material in enclosed structures autonomously, named RAD Copter.

After observing the systems, we conducted in-depth interviews with team leaders for each of the four assessed systems using the core questions for each criterion, as described in Section 3, to guide the discussion of each system. This method allowed us to apply the Schema in collaboration with team leaders.

A table providing a visual representation of the complete results of the assessments is contained in Appendix A. In sum, before we discuss specific aspects in-depth, all four systems we assessed qualified as robotic systems, i.e., they objectively all had sensing, computing, effecting, and communication capacities, at least to a minimal degree. All four systems evaluated also presented one or multiple dimensions of autonomy that were rated as "high" and, therefore, warranted full evaluation. Only one of the systems, Mobula, the fixed-wing UAV interceptor armed with a kinetic "net-gun," was determined to be designed with capabilities for which its intended function is to cause harm, triggering the evaluation of the eight additional criteria. However, it was clear that if misused, the other three systems could cause harm. This fact raises the question of whether a distinction should be made between systems designed with capabilities intended to cause harm and systems with capabilities that can cause harm, which we will discuss in greater detail in Section 5. However, for this study, we evaluated the three systems that were not designed with harming capabilities according to the twenty-nine standard aspects.

To illustrate the practical application of the Schema, one of the four systems we assessed shall now be discussed in more detail, namely the system called ANYmal. ANYmal is a four-legged robotic system. The specific model we assessed is one component of the CERBERUS Team (CollaborativE walking & flying RoBots for autonomous ExploRation in Underground Settings), which won the USD 2 million first-place prize in the DARPA subterranean (SubT) challenge Systems Competition in September 2021 [49,50]. DARPA's SubT challenge is predicated on seeking "novel approaches to rapidly map, navigate, and search underground environments during time-sensitive combat operations or disaster response scenarios" [51]. Eight teams competed for USD 3.5 million in prize money in the SubT Challenge [52] organized by DARPA, the mission of which to make pivotal investments in breakthrough technologies for US national security [53].

Of the four systems we evaluated in our study, ANYmal was the most advanced. The specific model we looked at was a version of the commercially available ANYmal C platform developed by ANYbotics [54]. Of particular note, the ANYmal we evaluated is one component of a multi-modal perception system, including both walking and flying systems, with self-organized, networked communications designed to optimize efforts to search unmapped, subterranean environments. In other words, the ANYmal we analyzed was part of a system of systems.

ANYmal is capable of fully autonomous operation for approximately one hour before needing to return to a battery charging station. It is designed to "explore" with no human input, perceiving and maneuvering through unknown environments. ANYmal can also adapt its behavior based on previous experience for training purposes, but it is not in a "learning" modality during operation. Given these attributes, ANYmal is considered to have autonomous capabilities according to our evaluation tool. Additionally, ANYmal is not intended to cause kinetic, sensory, psychological, or other harm but rather perceive its environment without affecting it. Given these attributes, 29 standard criteria were analyzed according to our evaluation tool.

Our tool identified some specific properties of the ANYmal as potentially posing ethical concerns. The criteria Appearance and Cybersecurity were not considered in the design process (these aspects were neglected in the design of the other three systems too). For the case of ANYmal, we found that the defined conditions of the SubT Challenge resulted in the team deprioritizing certain considerations. For example, although the wireless communication system that ANYmal utilizes is vulnerable to hacking, the threat is considered low; therefore, communication encryption technology was not used. Similarly, given the controlled nature of the SubT challenge, the team had not considered the appearance of the ANYmal.

Given the stated purpose of the ANYmal, our finding is not surprising. However, suppose the next step in ANYmal's development was deployment in actual search and rescue operations, which the SubT Challenge is designed to mimic. In that case, we must ask when will appearance and cybersecurity considerations be addressed? Furthermore, if such considerations are not an inherent part of the design process, there is the possibility that they will be more challenging to address in the long term. Given the inherent vulnerability of people trapped during an emergency search and rescue operation, all potential sources of unintended harm must be deliberated. Consider, for instance, the perspective of a person being approached by a robot for the first time in their lives when trapped in near-darkness beneath rubble—would its color, motion, and overall effect result in greater fear or feelings of relief?

Our assessment also identified Emergent Properties as an aspect of the ANYmal that presented a potential risk of ethical concerns. The core question that guides the evaluation of this criteria is as follows: "Can system-to-system interaction yield unexpected or emergent properties?" As part of a framework of several systems designed to conduct collaborative exploration and mapping through the merging of data, emergent behaviors are expected. In a controlled environment, such as the DARPA SubT challenge, emergent behaviors are not a serious concern. If, however, a system is deployed in a real-world environment, behaviors that emerge as a result of a system's interactions with other systems could have ethical implications, especially in sensitive operations such as search and rescue missions, when human lives are on the line.

Two other aspects of the ANYmal that our evaluation identified as potentially raising concerns were Dual Use Management and Misuse Prevention. These two related but distinct criteria also exemplify the nuance that must go into a system evaluation and that such an evaluation applies to a specific case in time.

The core question guiding the analysis of Dual Use Management is as follows: "Are risks of dual use explicitly expressed and—if possible—have physical means to reduce the risk for dual use been taken into account?" Our analysis resulted in an amber rating and medium weight as dual use was not addressed by the CERBERUS team and is possible. However, a red rating was unwarranted as ANYmal was not built so that weapons could be easily integrated into the system architecture.

In the case of Misuse Prevention, the core question is as follows: "To what extent is the system designed to prevent or make difficult unwanted uses or undue extensions of its scope of use?" As stated above, ANYmal is not armed and is designed to avoid obstacles when operating. However, there is nothing inherent in the design that would prevent ANYmal from being used to search for enemy combatants in a subterranean environment rather than its stated purpose of searching for victims in a natural disaster. These facts resulted in a red rating and medium weight as unwanted uses or undesirable extensions of the scope of the system's use are likely to occur and have not been addressed in the system's design. This finding confirms the intuition that it is not a great leap for a system such as ANYmal, with autonomous capabilities designed for emergency search and rescue operations, to be used outside the scope of their design. Given the powerful capabilities an unarmed autonomous robotic system such as ANYmal could offer a military force, our evaluation tool guides the engineers, roboticists, and programmers involved to be cognizant of the potential for misuse. Intentionality then results in the following question: Could a design feature be incorporated that prevents ANYmal from being used in combat operations?

5. Experience and Lessons Learned

By utilizing the assessment of specific autonomous security systems, we gained experience and learned lessons that may be relevant beyond the immediate context of our tool. They shall be discussed in this section.

The importance of addressing the notion of dual-use technology emerged as a major issue in our study. Robotic systems capable of autonomous functions create new use cases, which may not be what they were designed or intended to do. We found that the engineers, programmers, and roboticists that we spoke with are counting on government officials and international bodies to establish laws bounding their creations to "lawful" behavior. The individuals involved in developing these technologies were not well versed in the state of policy related to autonomous robotic systems. Given the lack of agreement on norms and regulations at the international level about autonomous systems, it is not a matter of willful ignorance of technology developers but rather a void created, intentionally or unintentionally, by policymakers.

Compounding the problem of a lack of agreement on international policies, roboticists are not necessarily aware of the real-world dual-use potential of their creations. The developers we spoke with were not well suited for analyzing how an autonomous system designed for a humanitarian emergency may have capabilities that are directly transferrable to a conflict zone. However, seemingly benign capabilities such as those designed for robotic subterranean search and rescue operations could be used to clear buildings in close-quarters combat or search and neutralize missions. Overall, we found that applying our tool to specific autonomous systems in dialogue with roboticists served to discover ethical hotspots and raise awareness on the part of roboticists. Hence, the discussion itself performs a critical function. An ethical assessment in practice is not only about a result neatly captured in a table.

In all four cases, aspects such as "Cybersecurity" and "Appearance" were not prioritized by the development teams. Interviewees stated that such aspects would be addressed in the future. In the case of government-funded projects in which research, development, and deployment phases are iterative and often fused, the question then becomes "when should a system be evaluated?" For an Article 36 Weapons Review, the answer is clear: A government must assess and clear a weapon before it can be legally used; if the weapon is modified, it must be re-assessed. This clarity does not exist for an ethical evaluation. When a system is not considered a weapon, even if it may impact human life through its operation, there is no point at which an ethical evaluation is required. As highlighted in Section 4, the potential for a system to cause harm, even if it is not designed as a weapon, also raises a question about our approach, namely, should a distinction be made between systems that are and are not designed with capabilities for which its primary function is to cause harm? It is clear that all of the systems we looked at could cause harm if misused; therefore, it may prove valuable to evaluate all criteria in all cases.

In reality, assessors likely evaluate prototypes at an early stage of development. In our study, all four systems evaluated were early prototypes. This presents both a limitation and an opportunity. By discussing the ethical implications of various aspects of systems with the teams responsible for research and design, we have the opportunity to influence the development process. On the downside, a prototype is hard to evaluate definitively as, by definition, it is subject to modification. The lack of a clear finishing line reinforces our impression that the impact an ethical evaluation tool such as ours can make stems from the questions it raises for engineers, roboticists, and programmers designing systems. They may then become more aware of ethical concerns and adopt a long-term vision of their processes. This benefit may be the most important, if not measurable, outcome of ethical assessments. This finding also implies that one should not expect to identify a universal code of machine ethics or that "global, socially acceptable principles for machine ethics" [55] (p. 59) can be easily elucidated through the assessment of specific systems (or otherwise). One widely cited project that emphasized the importance of public discourse on ethical issues that arise when operationalizing systems with autonomous capabilities is the MIT Media Lab's interactive project "Moral Machine" [56]. However, the "Moral Machine Experiment" was also criticized for problems presented by some of its more far-reaching statements [57]. We found that the act of analyzing specific properties of systems spurs practical conversations about what may be problematic and what can be performed to mitigate substantial risks. Our experience leads us to believe that the act of discussing actual systems, their properties, and the ethical and moral risks they present is vital regardless of whether or not clear-cut answers are identified.

Our assessment also made it clear that, despite an overlap between AI and autonomous robotic technology, our evaluation tool is limited. It is only applicable to autonomous systems that physically exist as embodied hardware systems. The tool is not easily transferrable to autonomous systems that exist as pure software or as systems of systems [58]. This limitation has far-reaching implications because embodied systems can often not neatly be separated from disembodied systems. Embodied systems are also running software, e.g., the operating system. That software may not be locally stored in the robot but elsewhere and in another "body." Nevertheless, it must be factored into a comprehensive assessment on equal footing with all other essential parts of a functioning system to be assessed, while pure cyber components cannot be adequately factored in. This situation begs the question of where the autonomous system to be evaluated begins and ends.

Moreover, there is also an IHL dimension to this problem that complicates matters further. The physical embodied space and the disembodied "cyber" space blend in today's armed conflicts [59,60]. Cyberattacks complement physical, kinetic attacks. The laws applicable to these two domains intersect, but the law of cyberspace is less developed than the established law of armed conflict [61]. Hence, when the two intersect in systems and their environment, systematic assessments become harder, rendering tools such as ours blurry at the edges.

Focusing on physically existing systems presents its own set of unique challenges. One challenge we ran into while conducting our research was how to evaluate inherently composite systems holistically. Each of the four robotic systems we analyzed consisted of an integrated collection of sensors, processors, actuators, and effectors. Generally, each component consists of its own collection of parts integrated into an element that is then combined into the whole. Some components may have autonomous capabilities while others do not. The question then arises as to how a composite system with autonomous aspects can be evaluated as a whole.

The composite nature of systems is further complicated because components of a robotic system can be modified or changed, redefining the system's capabilities and related ethical issues. For example, a system capable of autonomous flight with no effectors could be modified to include a human-operated kinetic effector (human in the loop), thereby changing the overall system and its ethical implications without changing any of the autonomous functions of the system. In his book *Wired for War*, Peter Singer addresses this possibility, stating, "You can integrate existing componentry and create a revolutionary capability" [62] (p. 31).

Furthermore, our assessment of specific autonomous systems made us aware of the links of our research with risk identification and management techniques. The use of risk-based approaches to practically address areas of concern of AI and autonomous systems is gathering momentum across the spectrum of governmental, private, and public discourse. For example, Microsoft has adopted a risk-based approach in addressing concerns about the technology it is developing. Microsoft's "harms modeling" evaluates (1) severity, (2) scale, (3) probability, and (4) frequency for each type of harm evaluated [63]. The US military also uses well-developed risk management tools to identify hazards, associated risk levels, control mechanisms, implementation methods, and residual risk levels [64].

The EU's proposal for a regulation on AI relies on the management of risks AI presents [32] (Article 9). It is possible that a framework such as Microsoft's Harms Modeling or a version of the tool described in this paper will need to be formally adopted and recognized as a standard method of assessing risks for autonomous systems.

The relevance of a risk-based perspective prompts us to think about improving and refining our evaluation tool further, beyond the experience we described above. Future research must entail further system evaluations. For the next iteration, we will explore modifying the weight value to better account for probability or the likelihood that a system aspect will result in inadvertent harm. This modification will bring the composite value of each evaluated property, i.e., each "criterion," into alignment with traditional risk matrices, identifying the severity of a problematic aspect and the probability that the aspect will result in harm. By aligning the tool with standard risk matrices, we will be able to share system evaluations in a more familiar format.

6. Conclusions

Through the discussion of our research to develop a practical tool that can be used to analyze the ethical issues presented by autonomous robotic systems, we hope to contribute to a conversation that goes beyond definitions and principles. By applying the evaluation tool described in this paper, our study generated knowledge about how lawyers, ethicists, engineers, roboticists, and operators can work together in the design process to develop systems that are built through informed discussion and clear intentions. In the absence of government regulations or international agreement on what is unacceptable, it is incumbent on the community of people working to develop new technologies to consider the impact of the technology on society. Given the complexity of autonomous systems, ethical considerations should not be ad hoc but wholistic, systematic, and informed by diverse perspectives. The evaluation tool we presented in this paper is not a final solution, but we hope it can spur an urgent conversation.

Our evaluation tool is not unique, but what distinguishes our research from others is that we have moved on from developing a theoretical tool to applying it to existing systems, with results that can be visually represented as in Appendix A. Real-world practical application has limitations, as described in Section 3, and brings unique challenges, as described in Section 5, but it also yields results and experience on a different level. However, it turned out that the critical challenge in applying an ethical evaluation tool resides not so much in the system itself or those designing it but in the accessibility of systems. In the domain of security, autonomous systems on the cutting edge of technological developments are shielded from the eyes of the public and researchers. Those who develop such systems are reluctant to expose and discuss the details of such systems. Hence, a significant amount of time and energy had to be invested into gaining access to systems. With the positive reception of our work and sustained efforts, we hope to access additional, cutting-edge systems so that our research can continue to progress and contribute to the urgent conversation concerning the development and use of autonomous robotic systems.

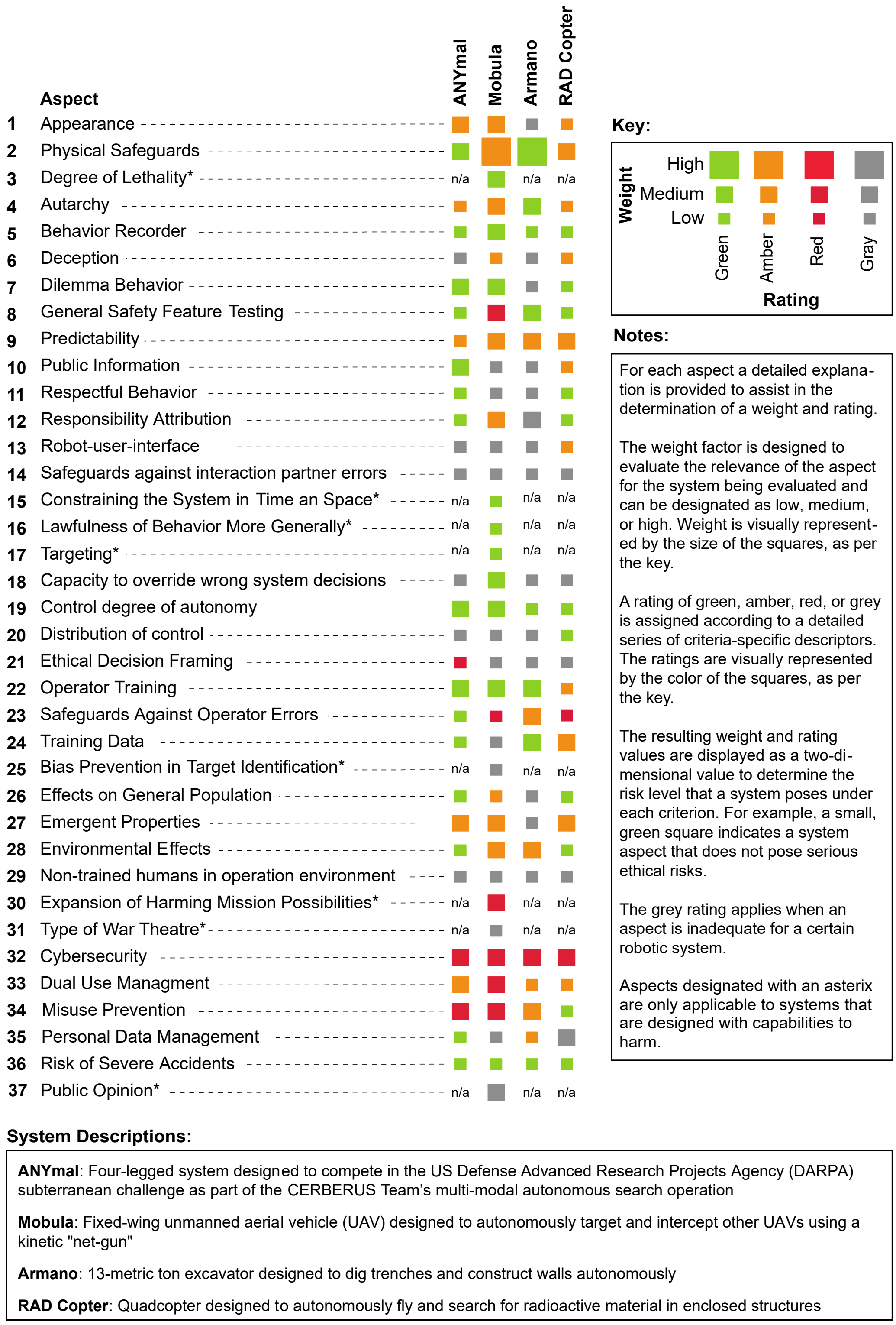
Author Contributions: Writing—original draft preparation, D.T. and T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Armasuisse Science and Technology and the Swiss Center for Drones and Robotics.

Acknowledgments: We are grateful to the researchers who provided access to their robotic systems and took time to discuss details of those systems with us and to David Danks for valuable feedback on a draft version of this paper. We also thank the two anonymous peer reviewers for taking the time to read the manuscript and for providing helpful comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A



	Aspect	ANYmal	Mobula	Armano	RAD Copter
1	Appearance	🟧	🟧	◼	🟧
2	Physical Safeguards	🟩	🟧	🟩	🟧
3	Degree of Lethality*	n/a	🟩	n/a	n/a
4	Autarchy	🟧	🟧	🟩	🟧
5	Behavior Recorder	🟩	🟩	🟩	🟩
6	Deception	◼	🟧	◼	🟧
7	Dilemma Behavior	🟩	🟩	◼	🟩
8	General Safety Feature Testing	🟩	🟥	🟩	🟩
9	Predictability	🟧	🟧	🟧	🟧
10	Public Information	🟩	◼	◼	🟧
11	Respectful Behavior	🟩	◼	◼	🟩
12	Responsibility Attribution	🟩	🟧	◼	🟩
13	Robot-user-interface	◼	◼	◼	🟧
14	Safeguards against interaction partner errors	◼	◼	◼	◼
15	Constraining the System in Time an Space*	n/a	🟩	n/a	n/a
16	Lawfulness of Behavior More Generally*	n/a	🟩	n/a	n/a
17	Targeting*	n/a	🟩	n/a	n/a
18	Capacity to override wrong system decisions	◼	🟩	◼	◼
19	Control degree of autonomy	🟩	🟩	🟩	🟩
20	Distribution of control	◼	◼	◼	🟩
21	Ethical Decision Framing	🟥	◼	◼	◼
22	Operator Training	🟩	🟩	🟩	🟧
23	Safeguards Against Operator Errors	🟩	🟥	🟧	🟥
24	Training Data	🟩	◼	🟩	🟧
25	Bias Prevention in Target Identification*	n/a	◼	n/a	n/a
26	Effects on General Population	🟩	🟧	◼	🟩
27	Emergent Properties	🟧	🟧	◼	🟧
28	Environmental Effects	🟩	🟧	🟧	🟩
29	Non-trained humans in operation environment	◼	◼	◼	◼
30	Expansion of Harming Mission Possibilities*	n/a	🟥	n/a	n/a
31	Type of War Theatre*	n/a	◼	n/a	n/a
32	Cybersecurity	🟥	🟥	🟥	🟥
33	Dual Use Managment	🟧	🟥	🟧	🟧
34	Misuse Prevention	🟥	🟥	🟧	🟩
35	Personal Data Management	🟩	◼	🟧	◼
36	Risk of Severe Accidents	🟩	🟩	🟩	🟩
37	Public Opinion*	n/a	◼	n/a	n/a

Key:

Weight	Green	Amber	Red	Gray
High	🟩	🟧	🟥	◼
Medium	🟩	🟧	🟥	◼
Low	🟩	🟧	🟥	◼
	Rating			

Notes:

For each aspect a detailed explanation is provided to assist in the determination of a weight and rating.

The weight factor is designed to evaluate the relevance of the aspect for the system being evaluated and can be designated as low, medium, or high. Weight is visually represented by the size of the squares, as per the key.

A rating of green, amber, red, or grey is assigned according to a detailed series of criteria-specific descriptors. The ratings are visually represented by the color of the squares, as per the key.

The resulting weight and rating values are displayed as a two-dimensional value to determine the risk level that a system poses under each criterion. For example, a small, green square indicates a system aspect that does not pose serious ethical risks.

The grey rating applies when an aspect is inadequate for a certain robotic system.

Aspects designated with an asterix are only applicable to systems that are designed with capabilities to harm.

System Descriptions:

ANYmal: Four-legged system designed to compete in the US Defense Advanced Research Projects Agency (DARPA) subterranean challenge as part of the CERBERUS Team’s multi-modal autonomous search operation

Mobula: Fixed-wing unmanned aerial vehicle (UAV) designed to autonomously target and intercept other UAVs using a kinetic "net-gun"

Armano: 13-metric ton excavator designed to dig trenches and construct walls autonomously

RAD Copter: Quadcopter designed to autonomously fly and search for radioactive material in enclosed structures

Figure A1. Overview of the Assessments of Four Autonomous Robotic Systems.

References and Notes

1. Majumdar, R.C.L.; Aoun, A.; Badawy, D.; de Alburquerque Bacardit, L.A.; Majane, Y.; Wilkinson, A. *Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)*; United Nations: New York, NY, USA, 2021.
2. Johansson, L. Ethical aspects of military maritime and aerial autonomous systems. *J. Mil. Ethics* **2018**, *17*, 140–155. [CrossRef]
3. Sparrow, R. Robots and respect: Assessing the case against autonomous weapon systems. *Ethics Int. Aff.* **2016**, *30*, 93–116. [CrossRef]
4. Yokokohji, Y. The use of robots to respond to nuclear accidents: Applying the lessons of the past to the fukushima daiichi nuclear power station. *Annu. Rev. Control Robot. Auton. Syst.* **2021**, *4*, 681–710. [CrossRef]
5. Christen, M.; Burri, T.; Chapa, J.; Salvi, R.; Santoni de Sio, F.; Sullins, J. *An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications*; University of Zurich Digital Society Initiative White Paper Series; University of Zurich: Zürich, Switzerland, 2017.
6. *U.S. Department of Defense Directive 3000.09. 2012; 8 May 2017*; Washington Headquarters Services: Washington, DC, USA, 2017. Available online: https://irp.fas.org/doddir/dod/d3000_09.pdf (accessed on 1 November 2020).
7. Group of Governmental Experts on Emerging Technologies and in the Area of Lethal Autonomous Weapons System. *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems; Convention on Certain Conventional Weapons; Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System*; United Nations: Geneva, Switzerland, 2019. Available online: https://undocs.org/en/CCW/GGE.1/2019/3 (accessed on 1 July 2021).
8. Scharre, P. *Army of None: Autonomous Weapons and the Future of War*; W. W. Norton & Company: New York, NY, USA, 2018.
9. Human Rights Watch and Harvard Law School International Human Rights Clinic (IHRC). New Weapons, Proven Precedent: Elements of and Models for a Treaty on Killer Robots. Available online: https://www.hrw.org/report/2020/10/20/new-weapons-proven-precedent/elements-and-models-treaty-killer-robots (accessed on 20 October 2020).
10. Future of Life Institute. Autonomous Weapons: An Open Letter from AI & Robotics Researchers. Available online: https://futureoflife.org/%20open-letter-autonomous-weapons/ (accessed on 15 May 2021).
11. CCW Protocol IV on Blinding Laser Weapons, adopted October 13, 1995, entered into force July 30, 1998, art. 1.
12. Campaign to Stop Killer Robots. Robots, C.T.S.K. About Us. Available online: https://www.stopkillerrobots.org/about-us/ (accessed on 15 September 2021).
13. United Nations Institute for Disarmament Research (UNIDIR). *The Weaponization of Increasingly Autonomous Technologies: Concerns Characteristics and Definitional Approaches, a Primer*; United Nations Institute for Disarmament Research (UNIDIR): Geneva, Switzerland, 2017. Available online: https://www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber (accessed on 15 May 2021).
14. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted June 8, 1977, 1125 U.N.T.S. 3, entered into force December 7, 1978, art. 1(2).
15. Human Rights Watch and Harvard Law School International Human Rights Clinic (IHRC). Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition. Available online: https://www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition (accessed on 15 May 2021).
16. Human Rights Watch and Harvard Law School International Human Rights Clinic (IHRC). Mind the Gap: The Lack of Accountability for Killer Robots. Available online: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots (accessed on 15 May 2021).
17. Santoni de Sio, F.; Mecacci, G. Four Responsibility Gaps with Artificial Intelligence: Why They Matteer and How to Address Them. *Philos. Technol.* **2021**. [CrossRef]
18. U.S. Department of Defense. Available online: https://www.defense.gov/about/ (accessed on 15 September 2021).
19. Arkin, R.C. *Governing Lethal Behavior in Autonomous Robots*; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2009.
20. Pasquale, F. 'Machines set loose to slaughter': The dangerous rise of military AI. *Guardian* **2020**. Available online: https://www.theguardian.com/news/2020/oct/15/dangerous-rise-of-military-ai-drone-swarm-autonomous-weapons (accessed on 15 July 2021).
21. ICRC. *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol 1 of 1977*; International Committee of the Red Cross: Geneva, Switzerland, 2006.
22. Geiß, N.B.S.B.R. Present futures: Concluding reflections and open questions on autonomous weapon systems. In *Autonomous Weapons Systems: Law, Ethics, Policy*; Bhuta, N., Beck, S., Geiß, R., Liu, H., Kreß, C., Eds.; Cambridge University Press: Cambridge, UK, 2016; pp. 347–383.
23. Boulanin, V.; Verbruggen, M. *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*; Stockholm International Peace Research Institute: Stockholm, Sweden, 2017.
24. Poitras, R. Article 36 weapons reviews & autonomous weapons systems: Supporting an international review standard. In *American University International Law Review*; American University Law Review: Washington, DC, USA, 2018; Volume 34, pp. 465–496.
25. Burton, J.; Soare, S.R. Understanding the strategic implications of the weaponization of artificial intelligence. In Proceedings of the 11th International Conference on Cyber Conflict, Silent Battle, Tallinn, Estonia, 1–17 May 2019.
26. Airbus. *Future Combat Air System (FCAS)*; Airbus: Leiden, The Netherlands, 2020; Volume 2020.

27. Knight, W. The pentagon inches toward letting ai control weapons. *Wired* **2021**. Available online: https://www.wired.com/story/pentagon-inches-toward-letting-ai-control-weapons/ (accessed on 11 May 2021).
28. Maas, M. Innovation-proof global governance for military artificial intelligence? How I learned to stop worrying, and love the bot. *J. Int. Humanit. Leg. Stud.* **2019**, *10*, 129–157. [CrossRef]
29. Backstrom, A.; Henderson, I. New capabilities in warfare: An overview of contemporary technological developments and the associated legal and engineering issues in Article 36 weapons reviews. *Int. Rev. Red Cross* **2012**, *94*, 483–514. [CrossRef]
30. Gustafsson, I. *How Standards Rule the World: The Construction of a Global Control Regime*; Edward Elgar Publishing: Cheltenham, UK, 2020.
31. Ethics Commission (Federal Ministry of Transport and Digital Infrastructure), Automated and Connected Driving. Available online: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile (accessed on 1 July 2021).
32. European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*; European Commission: Brussels, Belgium, 2021.
33. Burri, T.; von Bothmer, F. The New EU Legislation on Artificial Intelligence: A Primer. Available online: https://ssrn.com/abstract=3831424 (accessed on 11 May 2021).
34. Veale, M.; Zuiderveen Borgesius, F. Demystifying the Draft EU Artificial Intelligence Act. *Comput. Law Rev. Int.* **2021**, *22*, 97–112.
35. Burri, T. The New Regulation of the European Union on Artificial Intelligence: Fuzzy Ethics Diffuse into Domestic Law and Sideline International Law. Available online: https://ssrn.com/abstract=3865149 (accessed on 11 May 2021). [CrossRef]
36. Facebook AI: AI Now Enables Robots to Adapt Rapidly to Changing Real World Conditions. Available online: https://ai.facebook.com/blog/ai-now-enables-robots-to-adapt-rapidly-to-changing-real-world-conditions (accessed on 9 July 2021).
37. International Court of Justice, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion. ICJ Reports 1996, p. 226, 8 July 1996, para. 25.
38. International Court of Justice, Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, ICJ Reports 2004, p. 136, 9 July 2004, para. 106.
39. Arms Trade Treaty, adopted April 2, 2013, A/RES/67/234B, entered into force December 14, 2014, art. 7(1), UNTS I-52373.
40. Lustgarten, L. The arms trade treaty: Achievements, failings, future. *Int. Comp. Law Q.* **2015**, *64*, 569–600. [CrossRef]
41. Beauchamp, T.L.; Childreess, J.F. *Principles of Biomedical Ethics*, 8th ed.; Oxford University Press: New York, NY, USA, 2019.
42. van Wynsberghe, A. Designing robots for care: Care centered value-sensitive design. *Sci. Eng. Ethics* **2013**, *19*, 407–433. [CrossRef] [PubMed]
43. van Wynsberghe, A. Service robots, care ethics, and design. *Ethics Inf. Technol.* **2016**, *18*, 311–321. [CrossRef]
44. Asaro, P. What should we want from a robot ethic ? In *Ethics and Robots*; Capurro, R., Nagenborg, M., Eds.; IOS Press: Amsterdam, The Netherland, 2006.
45. Wang, N.; Christen, M.; Hunt, M.; Biller-Andorno, N. An Ethical Framework to Enhance Value Sensitivity in Humanitarian Innovation: Integrating Values in the Use of Drones in the Aid Sector. Forthcoming.
46. Roff, H.M.; Danks, D. "Trust but Verify": The difficulty of trusting autonomous weapons systems. *J. Mil. Ethics* **2018**, *17*, 2–20. [CrossRef]
47. Jobin, A.; Ienca, M.; Vayena, E. Artificial intelligence: The global landscape of ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
48. Jackson, P.T. *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*, 2nd ed.; Routledge: New York, NY, USA, 2016.
49. CERBERUS DARPA Subterranean Challenge. Available online: https://www.subt-cerberus.org/ (accessed on 1 May 2021).
50. September 24 2021 DARPA Press Release with the Official Subterranean Challenge Final Event results. Available online: https://www.darpa.mil/news-events/2021-09-24a (accessed on 25 September 2021).
51. Unearthing the Subterranean Environment: Subterranean Challenge. Available online: https://subtchallenge.com/ (accessed on 1 May 2021).
52. DARPA Subterranean Challenge Resources. Available online: https://subtchallenge.com/resources.html (accessed on 1 May 2021).
53. About DARPA. Available online: https://www.darpa.mil/about-us/about-darpa (accessed on 15 May 2021).
54. Autonomous Robots for Industrial Inspection. Available online: https://www.anybotics.com (accessed on 15 May 2021).
55. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The moral machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
56. The Moral Machine Project. Available online: https://www.moralmachine.net/ (accessed on 1 May 2021).
57. Kochupillai, M.; Lütge, C.; Poszler, F. Programming away human rights and responsibilities? "The Moral Machine Experiment" and the need for a more "humane" AV future. *Nanoethics* **2020**, *14*, 285–299. [CrossRef]
58. Trusilo, D.; Burri, T. Ethical artificial intelligence: An approach to evaluating disembodied autonomous systems. In *Autonomous Cyber Capabilities under International Law*; Liivoja, R., Väljataga, A., Eds.; NATO Cooperative Cyber Defense Centre of Excellence (CCDCOE): Tallinn, Estonia, 2021; pp. 51–66.

59. Innovation Board, AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Available online: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF (accessed on 15 September 2021).
60. Roff, H.M. Artificial intelligence: Power to the People. *Ethics Int. Aff.* **2019**, *33*, 127–140. [CrossRef]
61. Schmitt, M.N. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*; Schmitt, M.N., Ed.; Cambridge University Press: Cambridge, UK, 2017.
62. Singer, P.W. *Wired for War*; The Penguin Press: New York, NY, USA, 2009.
63. A Description of Microsoft's Approach. 2020. Available online: https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/ (accessed on 15 September 2021).
64. *Army Techniques Publication (ATP) 5–19 "Risk Management"*; Headquarters, Department of the Army: Washington, DC, USA. 2014. Available online: https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/atp5_19.pdf (accessed on 4 September 2021).