information

Article

An Empirical Examination of the Impact of Cross-Cultural Perspectives on Value Sensitive Design for Autonomous Systems

Austin Wyatt * and Jai Galliott

School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia; j.galliott@adfa.edu.au
* Correspondence: a.wyatt@adfa.edu.au

Abstract: The removal of direct human involvement from the decision to apply lethal force is at the core of the controversy surrounding autonomous weapon systems, as well as broader applications of artificial intelligence and related technologies to warfare. Far from purely a technical question of whether it is possible to remove soldiers from the 'pointy end' of combat, the emergence of autonomous weapon systems raises a range of serious ethical, legal, and practical challenges that remain largely unresolved by the international community. The international community has seized on the concept of 'meaningful human control'. Meeting this standard will require doctrinal and operational, as well as technical, responses at the design stage. This paper focuses on the latter, considering how value sensitive design could assist in ensuring that autonomous systems remain under the meaningful control of humans. However, this article will also challenge the tendency to assume a universalist perspective when discussing value sensitive design. By drawing on previously unpublished quantitative data, this paper will critically examine how perspectives of key ethical considerations, including conceptions of meaningful human control, differ among policymakers and scholars in the Asia Pacific. Based on this analysis, this paper calls for the development of a more culturally inclusive form of value sensitive design and puts forward the basis of an empirically-based normative framework for guiding designers of autonomous systems.

Keywords: values in design; autonomous weapon systems; military innovation; meaningful human control; cross-cultural comparison

Citation: Wyatt, A.; Galliott, J. An Empirical Examination of the Impact of Cross-Cultural Perspectives on Value Sensitive Design for Autonomous Systems. *Information* **2021**, *12*, 527. https://doi.org/10.3390/info12120527

Academic Editor: Luis Martínez López

Received: 4 November 2021
Accepted: 15 December 2021
Published: 17 December 2021

1. Introduction

The removal of direct human involvement from the decision to use lethal force raises a number of serious ethical and legal barriers that remain largely unresolved by the international community. Given the lack of human operators that could be held accountable for meeting the standards set by international humanitarian law, potential end-users must rely on alternative measures for ensuring accountability. Central to ensuring continuing accountability for these systems is maintaining meaningful human control, either through doctrinal tools or through deliberate design decisions. Broadly speaking, maintaining meaningful human control over future autonomous systems will require responses that can be grouped into two general categories. First, there will be solutions that target the organisational structures, norms and conceptions that shape how the system is used. In the case of ensuring meaningful human control over future autonomous weapon systems, an example of this kind of response would be the provision of specialised familiarisation training for officers tasked with overseeing deployments of weapon systems with some level of capability to operate autonomously. However, this article focuses on the second category, technical choices at the design stage, which shape the system itself toward more ethically acceptable behaviours and limit their capacity to be utilised for illegitimate purposes. Among the earliest suggested technical responses specific to autonomous weapons was Arkin's ethical governor model, which remains a useful example of the type of solution under this latter category. Essentially, these categories group responses on the basis of whether they achieve

an ethical outcome through shaping human behaviour and perceptions of the system, or by 'baking in' factors that promote ethical behaviour by the system. The concept of values in design is firmly in the latter category, in that it focuses on proactively embedding human values and ethical standards into the design of systems.

This paper focuses on presenting a consideration of how the value sensitive design methodology could assist policymakers, military planners, researchers and manufacturers in their efforts to develop increasingly autonomous weapon systems and military applications of artificial intelligence that remain ethically sound and under the meaningful control of humans. Through the empirical lens of a previously unpublished survey of policymakers and scholars in the Asia Pacific region, this paper will also challenge the tendency to assume a universalist perspective in the application of value sensitive design, and critically analyse how perspectives of key ethical and value-based factors differ across cultures in the region. Based on this analysis, this paper will present a call, supported by empirical, theoretical and documentary evidence, for the development of a more culturally inclusive form of value sensitive design that would be more amenable to the politically and ethically charged issue of lethal autonomous weapon systems in the Asia Pacific.

2. Review of Literature and Related Work

2.1. A Short Definitional Note on "Autonomous Weapon Systems" and Related Technologies

Before progressing to the main component of this article, it is worth taking a moment to clarify some of the underlying terminology that it will rely upon. One of the challenges in conducting empirical research examining issues related to autonomous weapon systems is that they continue to lack a universally agreed definition or even a set of common technical standards upon which one could base a determination of whether a given weapon system or platform would qualify as an autonomous weapon system.

This statement should not be taken as a suggestion that sterling efforts have not been made. The Group of Governmental Experts on Lethal Autonomous Weapon Systems (LAWS), operating under the auspices of the Convention on Certain Conventional Weapons, continues to work toward consensus, while non-governmental organisations, particularly the Campaign to Stop Killer Robots, and scholars have repeatedly called for binding regulations over their potential use in combat. Rather it is an acknowledgement that, while there is widespread agreement that designers should follow emerging principles (such as retaining meaningful human control), a consensus has not been reached on the specifics of what compliance looks like in practice. Indeed, it is this potential for end-users to hold discordant perceptions of ethical and legal principles that prompted this article to question the utility of the universalist approach at the core of values in design.

A review of the extant literature examining any of the multitude of ethical, legal or political issues stemming from the development of LAWS would illustrate to a reader the plethora of definitions that currently exist. Wyatt (2020) presents an explanatory overview for those with a deeper interest [1]. For this article, one can begin with the most used definitions: the United States military's definition and the loop categories promoted by Human Rights Watch. The former, contained in Directive 3000.09, essentially defines autonomous weapon systems as those that "can select and engage targets without further intervention by a human operator" [2]. Human Rights Watch presented a categorisation system, splitting relevant systems into one of three categories based on the level of human involvement in their Observe, Orient, Decide and Act (OODA) loop. Semi-autonomous, or human-in-the-loop, systems are human-activated with a limited capacity to autonomously maneuverer and/or engage designated target categories within geographic limitations [3]. Supervised or human on the loop systems can select and attack targets independent of human command yet include a mechanism that allows a human supervisor to interrupt or terminate the weapon's engagement process within a limited timeline. Finally, fully autonomous, human out of the loop, systems are capable of true autonomous function once activated by a human operator. While the specifics of this definition remain contested, for the purposes of this article it is sufficient to proceed with a slightly modified version of

the function-based definition adopted by Wyatt: "A fully autonomous Lethal Autonomous Weapon System (LAWS) is a weapon delivery platform, [informed by artificial intelligence], that is able to independently analyse its environment and make an active decision whether to fire without human supervision or guidance" [1].

2.2. Outlining Commonly Cited Approaches to Addressing Ethical Issues Raised by LAWS

The potential emergence of systems that remove, rather than merely distance, humans from the decision to employ lethal force brings with it a raft of ethical, legal and even strategic challenges, not least of which is the question of how to ensure accountability in warfare when there is no longer a direct causal link between the illegitimate use of force and a human who can be held to account in a courtroom. For some, the only solution to the issues raised by the spectre of "killer robots" is the institution of a pre-emptive ban under international humanitarian law on their development or use. While the prospect of a straightforward ban on the use of lethal autonomous weapon systems, modelled on the Ottawa Convention or the Protocol on Blinding Lasers, is theoretically a clear-cut and promising avenue for limiting the potential for harm, the past several years of stalled negotiations and the dual-use nature of the enabling technologies for increasingly autonomous systems suggest that this approach is becoming less feasible over time.

Short of a specific prohibition under international law (for example, through a binding treaty), the international community could pursue a "soft" or normative approach to regulating the use of autonomous weapon systems, central to which would be the establishment of common, if not totally universal, technical standards upon which common behavioural and de-escalation protocols could be built in the early stages of autonomous weapon system proliferation.

The principal argument in favour of a normative approach is premised on the belief that the informal nature of norms would allow technical experts to by-pass the political deadlock that attaches to efforts to formally define the limits of "autonomy". However, as Bode has noted that the flexibility of a normative-based approach to defining LAWS exacerbates the existing practice of states, as well as interest groups and manufacturers, of shifting how they describe a given system in order to promote it to potential end-users and buyers (by playing up its capacity to operate faster than humans outside of the need for direct control) or to shield a system from potential regulatory action (in which case they may highlight human supervision of critical functions or designate the system as highly automated rather than autonomous) [4]. This effect would then place greater importance on any potential discrepancies in how different state developers of autonomous systems, or even end-users, would apply even universally agreed values (such as the avoidance of harm to civilians).

A further benefit of this approach is the potential to build on pre-existing points of consensus in international humanitarian law, for example the prohibition of hostilities against civilians. Furthermore, the expansion of bilateral and multilateral military training exercises in the post-9/11 era, particularly among western militaries would suggest that military values may be far closer in their perceptions of ethical conduct of warfare than the lack of formal commitment could suggest.

The downside of this approach is that technical standards and normative behavioural frameworks lack the certainty of international law and would only guide rather than necessarily limit how states use autonomous weapon systems. This approach is, therefore, seen as insufficient at best (or a ploy at worst) by those who would prefer the implementation of a binding legal prohibition, and is thus unlikely to receive much attention in the official CCW negotiations.

Beyond questions of geopolitical and legal practicality, the core issue with the above-mentioned approaches is that they are concentrated on implementing state-led controls on how militaries conceptualise and deploy autonomous weapon systems, rather than targeting the design of the systems themselves. In essence, they are organisational-side solutions to issues raised by this disruptive innovation. In a sense focusing on the user-

side of the autonomous weapon system question is an understandable predisposition on the part of state and non-state experts involved in the Group of Governmental Experts negotiations. This is arguably the first time that the international community's capacity to prevent war crimes stemmed more from the technical characteristics of a weapon system than the training of its operators. While no weapon system could, or should, completely relieve those charged with authorising its use from accountability for the consequences of that deployment, this shift opens a third avenue for addressing the ethical and legal issues posed by autonomous weapon systems, building ethical compliance directly into the design process.

The concept of designing machines to behave ethically is certainly neither as novel nor as outlandish as it has sometimes been presented, particularly in the popular press. Notably, prior to the establishment of the GGE on LAWS, Ronald Arkin proposed that autonomous weapon systems should be programmed to undertake a series of gated decisions (based on simplified versions of the distinction and proportionality principles of international humanitarian law) before committing to the use of force, a feature that Arkin dubbed an "ethical governor" [5]. While there is a significant, and quite interesting, ongoing metaphysical debate over the possibility for an artificial intelligence to eventually develop sapience or its own internal morality, for the modern policymaker, it is unhelpful, and arguably unwise, to expect autonomous weapon systems to develop their own internal morality and act accordingly. Though it was worth distinguishing between morality, by which we mean following an internal, independent but societally influenced understanding of right and wrong, and ethics, which are externally imposed morally positive principles that mould one's behaviour toward the expectations of a community. As demonstrated by Arkin's ethical governor proposal, the latter is technically feasible, through the imposition of limitations on the guiding artificial intelligence that constrain the system from engaging in unethical actions. The often-cited concern with the ethical governor approach, however, is that current mechanisms for imposing limitations on a system that force it to behave in a manner that its designers believe is ethically sound are inflexible and often vulnerable to exploitation.

Unfortunately, the modern battlespace, particularly in the ground domain, is often a confused and rapidly evolving environment in which humans often behave in an unexpected manner. While a system may be able to recognise a Red Cross symbol, and thus refuse to engage that target, with a high degree of accuracy in laboratory conditions, there is no guarantee that it would be able to do so in a real battlespace. On a related note, such systems would also be highly vulnerable to perfidious actions by opposing forces, particularly in an urban insurgency. For example, an autonomous weapon system that was bound by a rigid ethical governor may be unable to recognise enemy combatants when their armaments are not in plain view or have been deliberately temporarily discarded [6], or a cunning foe could manipulate the system to accidentally engage targets that present a risk of civilian casualties that would have caused a human soldier to abort. Therefore, while imposing pre-set limitations or minimum criteria for engagement may be a workable solution for systems to be deployed in defensive maritime or aerial environments, it would not be a solution for ground-based systems.

2.3. What Is Value Sensitive Design?

Value sensitive design is a methodology that attempts to resolve these issues by injecting human ethics into an iterative design process. In essence, rather than assuming that technology is, by its nature, value-neutral, a value sensitive design approach sets off from the premise that human choices during and after the design of a technology imbue it with a set of values [7]. As a result, VSD asserts that a proactive integration of positive ethics into the design process is necessary and productive in limiting the possibilities for harm [7]. As a theory, this approach, therefore, sits at the intersection of the embodied values viewpoint [8] and safe-by-design paradigm [9]. Essentially, VSD is predicated on the premise that most misuses of technology can be minimised at the design stage [10],

a perspective that places a great deal of agency, and thus responsibility, in the hands of designers.

At the core of value sensitive design is its tripartite methodology. The first stage is a conceptual investigation. The purpose of this stage (whether it is completed initially or in subsequent iterations) is three-fold, to identify direct and indirect stakeholders, to analyse potential harms or benefits to each stakeholder group, and to identify which values would be connected to the technology [11]. The second stage consists of an empirical evaluation of potential stakeholder experiences and social impacts. The goal is to determine how stakeholders are likely to interact with the technology and determine what "prescriptions for, and restrictions on, action" [12] should be imposed [10]. The third stage is a technical [13] (or technological) [10] investigation, which consists of two steps. The first focuses on determining how the technology may support or constrain human values, while the second focuses on identifying and evaluating how the values identified in the earlier stages could be embedded in the design process [8]. This methodology is intended to be iterative and encourages conceiving of the stages as inter-dependent [13]. Nor are these stages always applied in the order presented here, sometimes one will start with the technical specifications of the system already known, in other cases the designer begins by identifying stakeholders and analysing potential harms. In some cases, for example, with a pre-existing technology where the author has no direct capacity to intervene in subsequent design evolution, it may be more useful to begin with the technical stage, identifying which aspects and functions of the technology harm or benefit values [10]. More important than the order of these investigatory steps is the participation of a multi-disciplinary team and a commitment to iterative experimentation.

2.4. Military Values and How They Are Also of Ethical Import

It is particularly important in this case to consider value sets particular to the military and the extent to which designers of AI and autonomous systems should prioritise or minimise them over more traditional civilian values. The most obvious example of the distinction between military and civilian values is in the taking of human life. Outside of the military context, the decision to inflict harm on another human is generally seen as antithetical to the values promoted by society. For a more specific example, consider that for an engineer overseeing the machine learning process for a program intended to govern a self-driving car, ensuring that the car deployed its brakes, or swerved depending on the speed, to avoid an unexpected pedestrian would almost certainly be the ethical design choice. The designers placed the preservation of human life at a higher ethical priority than delivering the passenger to their destination on time or preventing a minor car accident. In the military context, however, the designer must also consider military necessity. As a result, a designer may instead teach the AI alternative responses that reflect different priorities, such as pushing through neighbouring parked vehicles and continue its journey at speed in order to limit the risk of ambush, a suicide bomber, or an improvised explosive device. Granted, however, both scenarios would require that the system has sufficient autonomous capability, situational awareness and spatial capacity to undertake the 'ideal' response.

However, it is also incorrect and unhelpful to merely assume that military officers would dismiss the ethical risks posed by autonomous systems and always push for greater military utility in the design process. A 2020 study of trainee officers at the Australian Defence Force Academy found that reduction in the risk of harm to civilians was ranked as important at almost identically high levels as the reduction in the risk of harm to service personnel and that safety was the most concerning risk when asked to consider deploying into combat alongside autonomous weapon systems [14]. For comparison, reduced procurement and maintenance costs, the potential for deterioration of command authority and potential deterrent effect were all considered significantly less important by that cohort of future Australian military leaders [14].

Of far more value to this discussion is determining to what extent designers should account for potential harms and balance military values with civilian priorities in the development of potentially dual-use technologies. Repeated examples of platforms designed purely for civilian use being deployed in furtherance of armed strikes by state and non-state actors over the past several years have demonstrated that merely ignoring the potential applications of dual-use technologies is unsustainable, and continued debate over definitional minutia does not bode well for the efficacy of an international ban.

However, this is unlikely to happen while the international community, particularly within the academy, does not distinguish between those designers who agree to consider military values in their work and those who are actively developing weapon systems. The Campaign to Stop Killer Robots has led an effective effort to demonstrate the ethical challenges to research in this space and retains an important role in the ongoing UN discussions. Their calls were amplified by the highly publicised decision by Google employees to publicly oppose its participation in Project Maven, and the 2018 academic boycott of KAIST in response to a rumoured partnership with Hanwha Systems to conduct research toward autonomous weapon systems. The impact of these events is made clear in a recently published survey of attitudes among machine learning and artificial intelligence scholars, which found that 58% strongly opposed other researchers working on lethal autonomous weapon systems, 42% of whom would be willing to resign or threaten to resign over their university supporting such research [15].

These efforts can, and arguably have, had a chilling impact on the willingness of researchers to openly consider military values and potential military applications when conducting related research. The same study found that strong opposition was reduced by almost two-thirds, to 20%, when asked about developing AI tools for military surveillance and to only 6% when the AI application was limited to military logistics [16]. This distinction has some fascinating connotations for further research and would, theoretically allow for some level of military value consideration. However, given the current political climate around autonomous weapon systems, how can we expect junior and mid-career researchers to be willing to risk the kind of censure associated with accusations of promoting the development of "killer robots". Therefore, for value sensitive design to be a legitimate approach for maintaining meaningful human control in increasingly autonomous weapon systems, the international community must be willing to allow engineers and other researchers to openly consider military actors as legitimate stakeholders during each stage of the VSD methodology.

3. Methodology

A crucial first step in understanding how distinct value preferences may influence the implementation of value sensitive design for autonomous weapon systems, is to identify where those differences are. In order to contribute to this issue, this paper will draw on data obtained from a survey of 48 scholars and policymakers in the Asia Pacific, which was conducted in the first half of 2021. Delivered via the Qualtrics online platform, the survey attracted 48 responses from approximately 132 total academics and policymakers contacted (a rate of return around 37%). Participants in this study were all aged over 25 and located in eleven states (all bar two respondents resided in the Asia Pacific at the time of the survey). It is worth noting that participants were not uniformly distributed, this is unsurprising given that the majority of literature on this topic in English has originated in the United States or Europe. The largest groups were ten respondents from Australia and ten from the United States, with five choosing not to identify and 20 from respondents in ASEAN member states. All bar one of the respondents possessed a post-graduate degree (75% held doctoral degrees), and the gender ratio was 35 male, eight female, with five choosing not to identify.

The initial 132 academics that were approached to respond to this survey were chosen out of an initial group of scholars and policy makers identified in a snowballing reverse-literature review. This group was then evaluated to identify which individuals met one

or more of the following three criteria: first, evidence of recently published research on autonomous systems, international humanitarian law, or international relations within Southeast Asia; second, evidence of engagement with related public policy debate in the region; third, employment by a government department or tertiary education institution in the Asia Pacific. Contact was then made with these 132 academics and policy makers utilizing email, and followed ethical requirements set by the University of New South Wales Human Research Ethics Committee. Those respondents who expressed interest was supplied with further participant information and a link to the survey. We also promoted the survey through professional networks and social media platforms.

The survey instrument was developed with the advice of a cross-disciplinary and cross-cultural group of researchers. This voluntary instrument asked respondents to answer thirty-three questions, the majority of which were based around Likert scales or ranked responses. The survey was tested internally among the authors' professional networks before it was delivered, and the research team have subsequently received constructive feedback from both respondents and reviewers.

It is worth noting that the target audience for this survey was public policy, security studies and International Relations scholars and relevant policymakers with expertise in the Asia Pacific. While some of the contacted academics had published research related to autonomous weapon systems, the primary purpose was to understand the perspective of those in a position to influence policy in the region, rather than to conduct a survey of those specialising in AI, machine learning, or autonomous systems. Due to being comprised of experts in these fields, we must acknowledge that their responses should not be taken as representative of the general public. This was an intentional distinction, as there are already studies available that, at least somewhat, identify public opinion toward autonomous systems. The purpose of this research was to identify concerns among this group toward the potential for value sensitive design to influence autonomous weapon system development.

Identifying these concerns is of direct interest to any debate around the application of VSD to autonomous weapon systems because international relations scholars and policy experts would be well placed to both understand and inform the perspectives of government and military stakeholders in the design process. International relations and security studies academics, as a stakeholder group, are already contributing to the ethical and legal debate, particularly in the Group of Governmental Experts process. This is a group whose writings on key issues, such as operationalizing a definition of autonomy and conceptualizing meaningful human control, continues to have an outsized influence on the discourse. Furthermore, it is worth noting that the public has historically given greater weighting to the opinion of scholars and foreign policy remains largely the domain of policy makers and academics. The former trend was rapidly capitalized upon by groups opposed to autonomous weapon systems who touted academic, as well as industry, opposition in support of their position.

It is also important to flag up front the limitations of this survey. The first is the sample size; while the underlying study was initially envisaged as being delivered in person on a larger scale capitalising on academic conference attendance, the advent of COVID-19 related travel and event restrictions necessitated a scaling down. While the data collected still makes a valuable contribution to the literature, the smaller scale limits broad-scale generalisation of the results prior to a larger iteration in the future. Secondly, this survey is focused on the Asia Pacific, its intention is to prompt further regionalisation of the discussion of autonomous systems and military applications of artificial intelligence, which have to date been underreported in the literature. As this paper focuses on cross-cultural factors, the data is supplemented by previously published data with distinct geographic focuses, including the 2019 and 2020 IPSOS studies, the 2015 Open Robo-Ethics Institute survey, and the 2021 study conducted by Zhang et al. [15].

Despite these limitations, the data presented in this paper makes three key contributions to the literature and the ongoing debate around how best to limit the potential

for negative outcomes when it comes to military applications of AI and autonomous systems. The primary contribution of this data is that it is, to the knowledge of the authors, the most extensive survey conducted of scholars and policymakers in the Asia-Pacific, a region that (with the exception of China and the United States) has remained at the periphery of discussions around autonomous weapon systems. Secondly, the data allows this paper to present an empirically-based consideration of how distinct perceptions of values will have yet unconsidered, impacts on the applicability of value sensitive design to autonomous systems and AI. From this, one can draw the final major contribution of this paper, which is an examination of the flow-on effects of the export of systems designed solely on the basis of the order in which values are prioritised and perceived by designers and exporting states.

4. Results

4.1. Comparative Standards and Expectations of Transparency and Accountability

Although value sensitive design does not mandate that one start with a conceptual investigation, it is generally the logical place to start, particularly when one is looking to identify and evaluate differences in how distinct stakeholder groups prioritise values. To quickly reiterate, there are three goals of the conceptual investigation stage, to identify stakeholders, analyse potential risks and benefits to each group, and identify which values connect to the technology or issue under consideration.

The stakeholders associated with autonomous weapon systems present a particular challenge to the civilian researcher, as AWS are an innovation that, while reliant on dual-use technologies, remains largely under the purview of military laboratories, being designed principally for use by those in uniform, which can be an extremely difficult group to receive authorisation to formally interview, especially across multiple states. However, the military is no longer the isolated fiefdom that it was in the last century, and other groups will have a significant influence on how these systems are designed or whether they even see the light of day. As a result, the data focused on scholars and policymakers, which are stakeholder groups that give insight into how values around autonomous systems and military AI are likely to be perceived by actors and militaries in the region.

Although no state has admitted having fully developed a fully autonomous weapon system, much less outlined the values they followed in its design, the eleven guiding principles adopted by the GGE on LAWS, discussed above, can provide some guidance in identifying values. Reliability, accountability, human responsibility, security, transparency of process and function, and a clear delineation between civilian and military autonomous systems are all clearly apparent in the guiding principles.

We can also consider the flurry of activity in recent years in the closely related field of designing ethical military AI, which as the fundamental enabling technology for autonomous weapon systems, shed direct light on this issue. A recent evaluation of a selection of ethical AI principles conducted by Galliott, Cappuccio and Wyatt (forthcoming) [16] identified reliability, trustworthiness, safety, accountability, transparency, traceability, equity and value alignment, as commonly cited values in codes of AI ethics. As a more specific example, the US Department of Defense has publicly stated that military AI should be responsible, equitable, traceable, using transparent methodologies, data sources, and procedures. Galliott et al. [16] went on to propose a list of eleven principles for ethical military AI, among which these values were prominently featured. Their findings were based around a similar set of ethical principles to those considered in Umbrello's examination of data arising from the UK Select Committee on artificial intelligence (2019), which found transparency was the most commonly supported value, with accountability (sixth), responsibility (seventh), security (fourth) and privacy (third) all repeatedly being supported by evidence submitted to the committee from a range of stakeholders including academics, NGOs, governmental bodies and industry.

4.2. Cross-Cultural Comparison of Importance Placed on Accountability and Transparency in Autonomous Weapon Systems

While each of the values identified above would be of importance to an engineering team attempting to incorporate value sensitive design into their development of an autonomous weapon system, for the sake of brevity, this paper will focus principally on transparency and accountability to illustrate the level to which cultural paradigms can influence value expectations. The authors chose to focus on Transparency and Accountability partially due to their unique importance in ensuring value-positive autonomous systems and partially due to the fact that transparency and accountability are values that have been defined differently in different states and cultures. What we mean by this is that, with regard to governmental decision making, for example, Russian policy makers are able to make substantive policy decisions under far less public scrutiny and operate with much lower levels of transparency than their Australian counterparts. Russia is ranked 150 on the World Press Freedom Index (Australia is 25th), 137th of the Corruption Perception Index (Australia is 12th), and was scored at 20/100 by Freedom House in 2021 (Australia was given 97). Importantly, this is intended purely as an observation in support of the decision to focus on these values, and the authors have deliberately chosen not to place a moral value on either standard.

Ensuring that a system remains under the control of a responsible actor, and that actor can be reasonably held accountable for the system's actions, would be a necessary feature for autonomous systems to abide by the tenets of international humanitarian law and human conscience. In a 2019 study, the private research company IPSOS (on behalf of the Campaign to Stop Killer Robots) identified that 54% of those who stated that they opposed lethal autonomous weapon systems cited the belief that they would be "unaccountable" as a key factor in their opposition [17]. Despite Australia and the United States being the only two states that cross-over between the respondent groups in the IPSOS study and the underlying dataset for this paper, there are still some interesting takeaways. Principally, it is noteworthy that while this rate rose to 65% among Australian respondents, it fell to 36% among respondents from the United States. Ensuring accountability is especially important if one is to subscribe to the belief that engineers are capable of reducing or eliminating the potential for a system to be utilised in a manner that causes harm, as the absence of a direct human operator inherently raises the question of how to ensure systems are used in a legitimate manner, and to what extent designers and manufacturers should be held liable where a system is used in a negative or harmful manner.

In the relevant question, participants were presented with a scenario in which a hypothetical system operating within expected technical parameters (to rule out a pre-existing technical failure or manufacturing defect) experiences an unexpected error that leads to the system mistakenly engaging a civilian vehicle (the exact wording of the question was: "An autonomous system experiences an error while operating within manufacturer's specifications that affects its preprogrammed 'decision making' in a manner that leads to the system mistakenly engaging a civilian vehicle. Rank the following in terms of which group should bear responsibility"). Based on this scenario, respondents were asked to rank a series of stakeholders in terms of what level of responsibility they should be assigned for any resulting harm. The presented stakeholders were: The manufacturer as a corporation, the system's designer/architect, the military officer responsible for overseeing the system, the commander who ordered that the system be put into operational use, the politicians who allocated funding to procuring or developing the system, and the lawyers who completed the legal review of the weapon prior to its deployment. The goal of this question was to identify expectations of how accountability should be assigned in the event of illegitimate harm, which is among the core issues that continue to characterize the international debate. As a baseline, we can first consider the overall total rankings, which reflected a general perception of the most responsible stakeholders (if we combine responses that ranked them from first to third). Overall, the greatest percentage of respondents (56.7%) ranked the manufacturer as a corporation as highly responsible for the theoretical harm (24% first,

9% second, and 23.5% third). This is closely followed by the 56.3% who would blame the system's designers/architects (21% first, 14.7% second, and 20.6% third). That the most responsibility is placed on these groups is of particular importance given value sensitive design's contention that it was possible for engineers to eliminate the potential negative use cases of a technology. While the commander who authorised the deployment of the weapon system would be blamed by a smaller 47% of respondents, it is notable that the commander was held second most responsible for the theoretical harm at a significantly higher rate than either the manufacturer or the designers (23.5%), which would suggest that significant accountability should still be placed in the lap of the military officer who authorised forces to utilise the system, even where the manufacturer was believed to have not sufficiently limited the potential for harm. The other category of stakeholders were the politicians who allocated funds or otherwise signed off on the acquisition of the system and the lawyers responsible for conducting the Article 36 legal review, in that they were ranked sixth or seventh by 38.3% and 64.7% of respondents respectively. The overall impression from the data would suggest, therefore, that accountability is perceived to primarily sit with those who have direct input into the design of the system and the decision to place it in the situation where the error leading to harm occurred. In a way, this is similar to how military accountability is perceived with current generation weapon systems, where accountability of harm to civilians is generally assigned to the designers (in the case of a technical failure) or the system's operators (where the fault stemmed from the incorrect use of the system).

Breaking this data down on the basis of geography and strategic culture, there are some interesting takeaways that should be highlighted. The first is that, while the majority of respondents held those in the design and manufacture stage significantly responsible, this rate was even higher amongst Australian and the United States respondents, while it was placed in the second tier by the majority of Singaporean and Vietnamese respondents, and in the lower rankings by Indonesian and Philippines respondents. The gap is also apparent when respondents ranked the designers of the system's artificial intelligence, which was placed in the bottom three rankings by the majority of Singaporean respondents, but in the top rung by the majority of those in Australia and Vietnam, and in the middle by those in the United States and Indonesia. At the same time, however, no Australian respondents placed the most blame on the commander who ordered that the offending system be deployed, and only 28% ranked the commander second. Instead, the commander, as well as the directly supervising officer, were placed in the middle band by the majority of Australian respondents. Contrastingly, approximately two-thirds of Indonesian and Singaporean respondents placed the commander among the top-three most responsible for the error leading to harm, with similar rates among Philippine and Vietnamese respondents. Finally, while the majority of respondents placed only a minor level of blame on politicians and policymakers, there was more nuance on the individual state level. For example, 67% of Indonesian respondents and 75% of Singaporean respondents ranked politicians in the top-three most responsible stakeholders. Compare this to the 30% of Australian and 44% of United States respondents who ranked politicians similarly. There was a similar discrepancy in the number of respondents who placed some responsibility on military lawyers. Distinctions in the level of responsibility that policymakers would expect to be placed on different actors in the event of an unintentional or illegal engagement by an autonomous weapon system support the concern that different states may take distinct approaches to ensure accountability in those cases, which raises further concerns as to the continued legitimacy of future uses of force.

Central to ensuring accountability is the need for systems to be sufficiently transparent and recognisable that even those charged with auditing their use externally can understand how the system reached a particular decision or justified a given action. In essence, here we are talking about the "black box" problem (a system "for which we know the inputs and outputs but can't see the process by which it turns the former into the latter" [18]) and the challenges engineers face in designing systems to be secure and capable of operating

outside of direct human control while still ensuring that their actions remain intelligible, justifiable and auditable.

Certainly, the data collected in this survey supports the contention that developing and implementing effective transparency measures is vital for ensuring the ethicality of future deployments of autonomous weapon systems. When asked to consider this issue against a Likert scale, 72% of respondents considered military transparency to be an important aspect for any deployment of autonomous weapon systems to be ethical, and a further 21.9% considered it somewhat important. There was a similar level of importance placed on government transparency, with 69.7% believing it was important or somewhat important (18%) to ensure the ethics of future deployments. Given that the respondents were all educated scholars and policymakers, this result is unsurprising, and the only significant cultural variation appears in the percentage of respondents who were unsure. There were no respondents who thought government and military transparency were somewhat unimportant or unimportant.

Respondents were then presented with a list of six transparency factors and asked to rank them by the level of importance for ensuring that government and military use of autonomous systems remains ethical. The first factor was the *timeliness of public disclosure of when autonomous systems are deployed*, which was incidentally ranked most important at the highest rate (44.4%), with a further 29.6% ranking it second. The *consistency of government reporting* was the second most important factor for the majority of respondents (51.9%). While there was a broader spread of allocation among Australian and American respondents, there were no significant cultural variations between these two factors.

The next factor, *openness to NGO oversight and guidance*, was ranked fifth in terms of importance for ensuring transparency by 41% of respondents, with 30% placing it third and only 7.4% considering it the most important factor. On the individual state level, it is worth noting that Australian respondents (and one Belgian) were the only ones to consider this the most important, or second most important, factor. The majority ranking for this factor appears to have primarily stemmed from the Indonesian, Singaporean and American respondents, the majority of each of whom considered openness to NGOs as unimportant compared to the other factors for ensuring legitimacy. Overall, however, the fact that the significant majority of respondents considered openness to NGO oversight and guidance as of only middling to lessor importance to government transparency. This raises questions regarding calls for NGOs to take an active role in assessing or applying compliance measures for autonomous weapons, as well as the extent to which the Campaign to Stop Killer Robots should continue to have a significant role in the development of those future regulations.

When we combine this with the rankings that were given to the *involvement of NGOs in providing ethics training to soldiers*, one is left with concerns as to how much of a role NGOs could have in ensuring the transparent use of autonomous weapon systems. The fact that 77.8% of respondents considered this the least important factor (with a further 15% ranking it as the fifth most important factor) would suggest that respondents would not rely heavily on the non-government sector to teach ethical behaviours and promote transparency among soldiers charged with overseeing autonomous weapons. Arguably more importantly, however, it suggests that the seemingly logical basis for ensuring consistently applied values in design, which is the direct involvement of respected, external, and independent NGOs in training engineers and designers on military ethics, is not considered an important factor in measuring whether a principle one of those values is being met, being transparency. Given this result, it is particularly important for scholars to be pushing states to become first movers in the incorporation of NGOs into how they train their soldiers to operate alongside autonomous systems in a ethical manner.

The fifth factor was the *level of detail provided on actual systems adopted and deployed*, which most respondents firmly ranked in the middle tier (22.2% third and 48% fourth). Once we split the results based on geography, it is interesting to note that the only respondents who placed this as the most, or second-most, important factor were from the United States (only 22%), Vietnam, and Singapore. Australian, Indonesian, and Malaysian

respondents principally assigned this factor a middle-level ranking. Given the centrality of the question of what should be classified as an "autonomous weapon system" to the legitimacy and effectiveness of any future regulation, as well as the obvious challenges inherent in following the Russian suggestion that such determinations remain solely in the hands of individual states, it was surprising to see that outside the United States, Vietnam and Singapore, none of the respondents considered the public's capacity to access timely and accurate data on what systems the military is employing to be of significant importance in ensuring the use of those systems is sufficiently transparent. This raises obvious challenges to the concept of public accountability but provides some interesting guidance on how engineers could balance their response to the politically charged "black box" AI issue [19]. In the absence of effective public pressure driving individual states to require such disclosures from manufacturers and designers, the engineers that work in this space are not incentivized, and may even be dis-incentivized, from developing systems that allow for either external regulation or plain language statutory reporting.

The final factor presented to respondents was the *clarity of my military's controls on the use of autonomous systems*, which had the most even spread amongst total respondents in terms of the importance of ensuring that government and military applications of autonomous systems remain transparent. While there was not a significant geographic correlation for this factor, the transparency of the operational requirements that a military place on its decision to deploy a given weapon system is of less importance to the design stage.

Overall, this data supports the contention that the question of whether a universal set of ethics and values can be determined and enshrined in regulation is less important than a thorough understanding of how those values are perceived by those charged with implementing design choices. Consider firstly the example given in the accountability question, where an autonomous system experiences an unexpected error and illegitimately uses force against a human. While 'accountability' would arguably be the most likely value to be supported for inclusion in any future international regulation of autonomous weapon systems and will almost certainly continue to have pride of place in the CCW discussions, merely requiring its consideration at the design stage would not necessarily lead to its universal implementation. For a moment let us limit ourselves to the Australian and Indonesian respondents. The Australian respondents placed a heavier burden on the manufacturer and AI system designer than the average, while placing the commander who authorised its deployment and the officer charged with overseeing its use in the middle rankings. By contrast, the majority of Indonesian respondents placed the authorising commander in the top-three most responsible actors, and the manufacturer in the middle rankings. The question of which camp is correct is irrelevant for the purposes of this paper. What is interesting is how these diametrically opposed rankings would influence how an Indonesian design team and an Australian design team incorporated accountability into their methodology. Given this data, it would be defensible to argue that the Australian team would be more likely to focus its efforts on measures for ensuring manufacturer accountability (such as improved quality control, the dissemination of safety and technical standards, or the establishment of a government watchdog), while the Indonesian team's focus would be on measures within the military legal process or in the training and oversight of its commanders. Both teams would state that they have successfully incorporated accountability but would disagree with the other's approach, and neither would necessarily have met the intention of the hypothetical international regulation. Based on this data, we would expect to see a similar discrepancy if the teams were asked to embed transparency in the design of autonomous systems. For example, ensuring that the system allows for NGOs to provide guidance and feedback, was only considered to be in the top rankings by some Australian (and one Belgian) respondents, while the majority of respondents considered it far less important, would suggest that only the Australian design team would be likely to prioritise design choices that allowed for this involvement, particularly at the expense of other avenues.

However, we should also consider the possibility that these discrepancies are the result of distinct contexts and societal norms, rather than a disagreement on how to define a given value or the standard at which the respondents would believe that value was satisfied. For example, under this approach the difference in accountability levelled against manufacturers could be attributed to respondents from state A living in a society that places a higher legal burden on manufacturers, while state B's legal regime places greater responsibility on users in cases of harm arising from the use of a product. Even with this approach, however, the central problem is determining how pre-emplaced design constraints can allow a system to behave "ethically" when different environmental contexts would "shift the goalposts" so to speak.

However, in its entirety this data supports a conclusion that, for value sensitive design to be successfully incorporated into the design of autonomous systems and military applications of AI, it must account for the transitions and divergences that occur when geographically-, socially- and culturally-diverse groups translate the high-level value into guiding norms and regulations and then finally into value prioritisation and design choices.

5. Discussion

5.1. VSD as a Technical Design Stage Approach to Meaningful Human Control

A common argument in favour of pursuing a value sensitive design approach to AI and autonomous weapon systems is that part of prioritising ethics in the design process for these technologies would involve embedding meaningful human control elements into their very fabric. However, as with other proposals for making machines behave ethically, a significant problem with a VSD-based approach to autonomous systems and AI is that machines are inherently incapable of morality. Putting aside the concept of General AI, the type of functional, task-based artificial intelligence that is central to this debate does not possess sentience or free will. As they are incapable of forming their own morality, designers must settle for embedding parameters that ideally limit the system's potential actions to those within established limitations that correspond to, or are reflective of, what the designer would consider ethical boundaries. The realm of possible "ethical" behaviours for AI, particularly those relying on machine learning, is therefore limited to situations and decisions that were sufficiently serious, foreseeable and within the intended use case that they were identified as issues by the engineering team, and that a solution was technically feasible. Outside of these scenarios, current technology would limit the capacity of autonomous systems to behave "ethically" in an unexpected situation. This is further complicated by the fact that the system's designers can only account for situations that fall within their experiences and that they would consider sufficiently important that a given behaviour or restriction should be prioritised at the expense of another aspect of the project.

5.2. What Are the Problems with VSD in the Context of LAWS?

There are certainly well-reasoned arguments in favour of implementing value sensitive design as a tool for ensuring that future autonomous systems are limited in their potential negative impacts and remain under sufficient levels of meaningful human control regardless of operational and doctrinal decisions (which are generally seen as the domain of individual militaries). Unfortunately, this approach is not without its demerits, and we must now turn to those of particular relevance in the context of increasingly autonomous systems and military applications of artificial intelligence.

The first issue with relying on value sensitive design in this context is that it would require that one commits to the underlying assumption that "certain values are universally held, and that this fact can provide normative direction in design" [20]. While VSD also acknowledges that culture plays a role, it is limited to influencing how these values "play out" in a given time period, and although some values are culturally specific, there are also universally applicable factors upon which the methodology can rest [21]. Those limited values that can be said to be universally present across cultures are outnumbered

by those that are either absent or differently interpreted by different cultures, even within geographic regions. The problem, therefore, with a "universalist" approach has less to do with the existence of globally accepted values among those that are more relative, than it does with the risk that policy-makers and scholars will assume that engineers, designers or even end-users will act in a particular way [22], when in fact holding the same "value" on paper may translate into very different design decisions once it has passed through the lens of culturally distinct normative frameworks. This leads to another potential pitfall in the case of autonomous weapon systems.

Assuming a universally applicable set of values as the starting point for international regulation of an emerging technology presents an additional risk that actors will, whether innocently or maliciously, claim to have designed or procured systems that meet a given value (for example, accountability) but the standards used may be different from those of another state, or even the designer in the case of a purchased system. A related question is that when designing the underlying list of values to be considered by a development team, it becomes necessary to grapple with the question of who should decide which values should be considered and how should value prioritisation decisions be made. If those questions have different answers based on cultural, organisational, legal, doctrinal factors, for example, then we cannot be certain that even if the international community was to agree on a universal set of values that should guide autonomous system development, that it would result in universal design decisions.

Given the centrality identifying, analysing and prioritizing values to the concept of value sensitive design, it is particularly concerning that after seven years of formal discussions, there remains no agreement on the finer details of key concepts in this space [1]. Despite ostensibly being the principal venue for such discussions, the ongoing Group of Governmental Experts on LAWS process, organised under the auspices of the UN and the Convention on Certain Conventional Weapons, has yet to yield a universal definition of autonomous weapon systems, much less technical standards to guide engineering teams. This concern arises from the fact that an absence of commonly agreed definitions, technical standards or governance measures reflects the level of difficulty that is likely to characterize attempts to develop any universalist value set that could guide developers that wish to integrate positive values into autonomous systems.

The closest the process has come to reaching a normative framework was the adoption of the eleven Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, in September 2020 [23]. The most relevant of these principles to this paper are:

- "Human responsibility for decisions on the use of weapon systems must be maintained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system";
- "Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control";
- "Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems";
- "Discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies".

Although these principles are neither binding nor specific enough to build a domestic legislative framework upon, they do provide some valuable guidance for those interested in incorporating VSD into autonomous weapon systems. The first salient point to draw from these principles is that the GGE clearly places concern on maintaining a sufficient level of control over autonomous weapon systems that human accountability and responsibility can be maintained. That this requirement is reflected so strongly in the principles suggests that the GGE delegations, and the non-governmental representatives, remain

at least somewhat optimistic that a compromise can be reached short of a pre-emptive ban, which would require future systems to retain a sufficient level of meaningful human control, as well as in their capacity to generate an objectively assessable standard for determining whether a given system meets those standards. This is certainly an area where the incorporation of a value sensitive design methodology into autonomous system development and conceptualisation could have a significant impact on translating the GGE's guiding principles into a technically and diplomatically feasible set of regulations.

We can also see in these principles a commitment to the idea that risk assessment, open communication, and cross-disciplinary consideration of technical and policy challenges should be incorporated into every stage of the conceptualisation, design, testing and development cycle. From this, we can logically infer those considerations of risks, responses, and design decisions would also be expected in the processes by which militaries would select, procure, and incorporate these systems into their arsenals. The tripartite, iterative methodology of value sensitive design would appear to suit these requirements quite well, particularly with its insistence on stakeholder principles guiding, and in turn being shaped by, technically feasible design choices. However, these principles also highlight a problem with applying a universalist methodology in that it would raise the question of who should decide which stakeholders should, and should not, be heard in the process, and what weight should be given to each value in the design process. As well as, arguably more importantly, who should decide in which order values should be prioritised or that a state's declaration of considering values in the design of an autonomous system meets or fails some external standard for measuring the legitimacy of that process. That being said, however, VSD is conceptualized as an iterative methodology, which suggests that broader thinking is necessary. We could, for example, consider the potential interaction between VSD as a guiding approach that informs, and is informed by, governing bodies, which could make the approach more flexible and capable of evolving alongside shifting policy positions and values among principle-stakeholder groups.

Finally, the guiding principles are clear that the ongoing discussions of a ban under international law should not interfere with the development of civilian technologies. This is particularly problematic given the inherently dual-use nature of both robotics research and artificial intelligence training methods. The question of how to balance protecting the capacity of researchers and corporations to pursue legitimate civilian innovations in these spaces and the desire to limit the potential for autonomous weapon system development or proliferation is both complicated and unresolved. In the absence of some clear way of delineating between the two, additional restrictions on autonomous systems research and continued pressure on those working on research with potential military applications (but not direct weapon systems) will continue to have a chilling effect, particularly among junior researchers and corporate funders. Unfortunately, value sensitive design would not, by itself, resolve the issue, although incorporating transparency and cross-disciplinary cooperation into the research and design processes for AI-related research would be a valuable first step.

Although most state delegations at the most recent meeting of the GGE on LAWS (August 2021 at the time of writing) agreed to support these guiding principles, a closer examination of submissions to meetings of the group, as well as broader literature, is illustrative of deep divisions on the topic of autonomous weapon systems. Somewhat expectedly, the major division between state delegations is between those who have publicly supported a strong legal regulatory regime be established (of which a smaller number support a "ban" in the traditional sense), those who have agreed to participate in the CCW GGE process but stopped short of strong support for additional regulations, and those who have, for various reasons, pushed back on a ban, acted disruptively, or put forward support based on unworkable understandings of "autonomous weapon systems". In August 2020, Human Rights Watch [24] noted that, of the 97 countries that had publicly commented on the question of autonomous weapon systems, 30 states had expressed support for a ban. These states were Algeria, Argentina, Austria, Bolivia, Brazil, Chile,

China (issues with this inclusion are discussed later in this article), Colombia, Costa Rica, Cuba, Djibouti, Ecuador, El Salvador, Egypt, Ghana, Guatemala, the Holy See, Iraq, Jordan, Mexico, Morocco, Namibia, Nicaragua, Pakistan, Panama, Peru, the State of Palestine, Uganda, Venezuela, and Zimbabwe. Reaching Critical Will maintains a very helpful list of submissions to the GGE, in which countries have announced their position on a legal framework, some of whom have outright called for a developmental or deployment-based ban. States that took the view that a ban on fully autonomous weapons is necessary at the 2020 and 2021 meetings included Brazil, Chile, Mexico [25], Canada, Costa Rica [25], and Venezuela [26]. Somewhat unsurprisingly, none of these are traditional great power states, nor are they known developers of increasingly autonomous military technologies. A less cynical perspective would focus on the fact that their opposition to a ban has been couched in humanitarian terms, following the lead of the Vatican in opposing the transfer of the decision to end human life to a non-human entity on ethical as well as geostrategic grounds.

The second group, those states that have committed to supporting and participating in negotiations within the GGE without calling for a blanket ban on autonomous weapon systems, is also significant. Most support some level of additional legal or normative restrictive frameworks but stop short of supporting a pre-emptive blanket ban. For example, the Philippines delegation has called for a normative and operational framework and legally binding restrictions under the Convention on Certain Conventional Weapons but differentiates these calls from its support for a ban on anti-personnel LAWS [27]. Again, a failure to enthusiastically support calls for a blanket ban on "autonomous weapon systems" is unlikely to be due solely to either geostrategic or ethical concerns. For example, the Finnish delegation in 2021 argued that, while a pre-emptive ban could be imposed on LAWS "that operate in a truly independent manner or which perform tasks with no limitations and/or without a tasking by a human" [28], there remains a significant grey area with emerging technologies and the international community should focus on building processes for separating out legitimate applications so that innovation is not stymied by unrealistic oversight measures [28]. It is noteworthy that states, such as Australia, have instead put forward arguments in favour of more moderate control mechanisms and stated that discussion of "a prevention or ban treaty on LAWS is premature" [29]. It is also worth noting that the application of AI to military logistics have been generally deemed separate from weapon systems and targeting tools in the international discussions. While disagreements remain between states on the specifics of a definition for autonomous weapon systems, published definitions would not cover back-end integrations of AI in any pre-emptive ban. Finally, some states have put forward suggestions on control regimes that build on or incorporate existing compliance mechanisms, including France and Germany.

It is worth explicitly noting here that, despite appearing as a supporting group on the Campaign to Stop Killer Robots list, the Non-Aligned Movement's submissions to the GGE on LAWS are indicative of a more complex position. With more than 120 member states and a well-known commitment to non-interference, it is difficult to rely on these statements as comprehensively reflective of the individual member's positions. Indeed the 2018 statement from the Non-Aligned Movement referred to a re-emphasis of the position adopted at the XVII Summit of the NAM and 2018 NAM Ministerial Meeting (the Baku Declaration). However, the 206-page XVII Summit Final Declaration contains only a single paragraph that refers to LAWS, while the Baku Declaration does not actually mention autonomous weapon systems [30]. Taken together, it is difficult to agree with the Campaign that the NAM supports a ban rather than further discussions, and a closer examination of statements made by member states suggests that the NAM is a good case study of a moderate faction within the GGE debate.

The final "faction", to continue a metaphor, in the ongoing GGE on LAWS discussions are those, typically great power states, who have sought to oppose, disrupt, or block the group's deliberations, and publicly stated scepticism regarding the need and utility of a ban. Chief amongst these is, of course, the United States, which has an unfortunate

history of reluctance to compromise its pursuit of a military technology at the urging of the international community, examples of which include its continued refusal to sign the Ottawa Mine Ban Treaty, recent walk-back of a domestic policy disallowing the use of anti-personnel mines [31], and its refusal to ratify the UN Treaty on the Prohibition of Nuclear Weapons [32]. Given the centrality of military applications of artificial intelligence and increasingly autonomous weapon systems to the Third Offset Strategy, it is hardly surprising that the United States would not react with great support to efforts to introduce binding international legal restrictions on the perceived basis of their future force. For similar reasons, one may have been surprised by China's 2018 proclamation that it would support a pre-emptive ban on lethal autonomous weapon systems [1]. However, once one reads the submission carefully, it becomes clear that the Chinese delegation is proclaiming support for a ban on a surprisingly small category of system that would effectively overlook most systems currently proposed by states. For example, their definition of LAWS would discount systems intended principally for anti-vehicle or anti-material roles, as well as those that retain the capacity for any intervention by a human supervisor (including the capacity to manually abort or shut down the system during operations) [1]. Relying on this support would lead to a ban that would only cover a fraction of what opponents of autonomous weapon systems are actually proposing, and the argument that this was a genuine contribution in favour of the pro-ban camp must be considered sceptically by analysts, especially given the centrality of autonomous weapon systems and military AI to "intelligentized warfare" and the expectation among Chinese military planners that autonomous systems will be central to the next Revolution in Military Affairs [33]. The last great power that could be grouped into this final camp is Russia. Given its public and significant investments in military robotics in recent years, it is hardly surprising that the Russian Federation would also take a dim view of the proceedings, and the Russian delegation has repeatedly called for the GGE on LAWS to focus first on developing a universally agreeable definition of autonomous weapon systems, while simultaneously putting forward a definition that would exclude systems that retain some engagement with a human operator (the delegation explicitly supports excluding UAVs from the debate) [34], is limited to physical (rather than cyber) platforms, and that the final responsibility for assessing whether systems meet a definition of "autonomous" and are used responsibly should remain with the individual states [34]. Russia has also argued (as has the United States and others) that existing international law is adequate for regulating the use of autonomous weapon systems [35]. While these are genuine and important issues, when put into the context of Russia's pattern of disruptive and provocative behaviours in recent years, particularly in cyberspace, it suggests an intention to slow the CCW process before autonomous weapon systems become a reality. It is worth noting that each of these three states has made arguments in support of autonomous weapon systems on ethical grounds, arguing that the technology may be more effective than human operators in certain situations, may assist in the protection of civilians, and that they would lessen the physical and psychological toll on soldiers.

Granted, some of these divisions will have arisen from competing geostrategic or commercial interests; however, the continued debate over the finer details of virtually every term in the autonomous weapon system lexicon does not bode well for the argument that a universalist set of values could be codified. This is further complicated by the fact that the emergence of autonomous weapon systems is occurring in a competitive, multipolar series of development efforts. Therefore, while we can utilize VSD to identify differences between how individual actors are planning to implement meaningful human control and generate a better understanding of where common ground can be found, we cannot operate under the premise that there is a single design process upon which we can impose a VSD conceptual framework.

Implementing a globally accepted approach to value-sensitive autonomous systems would require a variety of trusted actors to represent different stakeholder groups (which creates further issues around how many global militaries and legal systems can be rep-

resented in a manageable stakeholder set), publicly verifiable access to in-development or proposed autonomous weapon systems (as well as the military officers guiding the development of accompanying strategic requirements), and universally agreed and sufficiently specific technical standards that could be applied. At the present time, however, attaining all three of these requirements appears impossible, which represents a serious barrier to implementing VSD on a sufficient scale with uniformity. This is particularly unfortunate given the fact that value sensitive design is built around the concept that one can identify, empirically analyse and prioritise certain values over others in the design of the given technology.

Furthermore, if the international community cannot show a sufficient level of uniformity on what values should apply to autonomous systems, as well as an understanding of how a given military would prioritise those values, then the approach will not be effective. While most advanced western militaries have spent the past two decades promoting the concept of interoperability and coalition operations, there remain differences even between close military allies around operational culture, doctrine, interpretations of international law, and legitimate uses of force. The role of distinct military and strategic cultures on how different states integrate a given technology into their arsenals is well documented in the military innovation literature. Arguably the most comprehensive recent publication on this issue is Adamsky's *The Culture of Military Innovation* [36], although other useful discussions have been written by Raska [37], Horowitz [38], Goldman and Leslie [39]. Therefore even if one was to limit the discussion to close military allies that share similar cultures and generally hold similar norms of military behaviour, for example, Australia, the United States and the United Kingdom, differences remain in how soldiers are trained to consider the use of force, the legal requirements for testing and adopting new weapon systems, and distinct geostrategic goals and operational taskings, that would complicate the roll out of an autonomous weapon system developed purely by the United States. Obviously, these challenges grow when one considers the impact of a system being developed by Russia or China, both of whom have well-documented track records of exporting advanced weapon systems to nations outside of their traditional sphere of influence.

The distinctions between how autonomous systems are viewed by states and the demonstrated market for remote operated uninhabited systems tie directly into a second problematic aspect of value sensitive design, its assumption that it is possible for engineering and design decisions to substantially, if not completely, remove the possibility that a system will be used in an unintentionally harmful or nefarious manner. Regardless of whether this assumption is generally defensible, which the authors believe it can be, it is particularly problematic in the case of autonomous weapon systems. This comes down to the fact that autonomous systems would lack the capacity to interpret contextual information and apply circumstance through interpretations of pre-programmed frameworks reflective of norms. In effect, the argument goes, the capacity for autonomous weapons to act in an "ethical" manner would depend solely on the capacity of designers to foresee potential scenarios and provide hard-coded guidance prior to deployment. This is further exacerbated by the absence of well-trained and responsible human operators, who would ideally be able to intervene in the event of failure.

Finally, we should consider the fact that, in the absence of universally agreed to technical standards or a common definition, there is an open question as to whether the quality standard deemed acceptable by designing engineers in one state would be accepted by those in another state. Opening the internal decision logic or giving access to core code may somewhat alleviate the impact of this disconnect; however, it would require militaries to be willing to allow external reviewers access to this sensitive data and could further open future autonomous systems to the risk of interference by malicious actors.

However, even if we put aside the question of the technical feasibility and diplomatic challenges involved in relying on a series of separate engineering teams to limit the potential harms of autonomous weapon systems, and accept the premise that engineers can prevent most misuses of technology and implement ethical uses through design choices, which is

a core assumption of value sensitive design, the question remains, whose values are we asking them to impose on autonomous weapon system design?

6. Conclusions

While the authors acknowledge that the preceding analysis is based on a comparatively small dataset, it is nonetheless indicative that there is sufficient divergence in how values are likely to be ranked during the value sensitive design methodology, based at least partially on the culture of the designers and policymakers. Given the fact that autonomous systems will, by definition, be more reliant on engineering and design decisions for ensuring their behaviour remains ethical, it is even more important to identify not only how to technically meet those standards but to understand how those standards and expectations shift across cultures.

With that said, the authors maintain that value sensitive design remains a methodology that has significant promise when it comes to identifying a feasible and practical mechanism for limiting the potential for harm stemming from autonomous weapon systems without resorting to a pre-emptive ban. However, the differences in how stakeholders would rank the above variables in terms of importance for meeting their standards of accountability and transparency suggests that the universalist mindset that continues to permeate the discourse around VSD needs to be altered. While the most straightforward solution would be to insist on some sort of supranational assessment and enforcement mechanism for ensuring that systems are correctly classified as autonomous and that the developer has meaningfully engaged with the value-sensitive design methodology, the fact remains that leading states (particularly the great powers) are unlikely to relinquish control over a technology that has been widely identified as vital to the future force paradigm. Instead, we suggest using the proliferation of cyber security ethics frameworks as a model, where centralised, broad values are identified and championed by industry and professional bodies, but individual designers work toward varying standards based on internal culture and the expectations of their end-users.

Author Contributions: Conceptualization, A.W. and J.G.; methodology, A.W. and J.G.; formal analysis, A.W.; investigation, A.W.; resources, A.W. and J.G.; data curation, A.W.; writing—original draft preparation, A.W.; writing—review and editing, A.W..; visualization, A.W.; project administration, J.G.; funding acquisition, A.W. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Australian Government through a grant by the Australian Department of Defence. The views expressed herein are those of the authors and are not necessarily those of the Australian Government or the Australian Department of Defence.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Human Research Ethics Committee of University of New South Wales (HC 200772; 11 November 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The full dataset referred to in this paper has not yet been publicly archived. Please contact a.wyatt@adfa.edu.au with any questions or requests.

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.

References

1. Wyatt, A. So Just What Is a Killer Robot? Detailing the Ongoing Debate around Defining Lethal Autonomous Weapon Systems. In *Wild Blue Yonder*; Air University Press, Maxwell Air Force Base: Montgomery, AL, USA, 2020.
2. Department of Defence, "Directive 3000.09". 2012. Available online: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf (accessed on 24 August 2021).

3. Righetti, L.; Sharkey, N.; Arkin, R.; Ansell, D.; Sassoli, M.; Heyns, C.; Asaro, P.; Lee, P. Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects. In Proceedings of the International Committee of the Red Cross, Geneva, Switzerland, 26–28 March 2014.
4. Bode, I.; Watts, T. *Meaning-Less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS*; Center for War Studies: Odense, Denmark, 2021.
5. Arkin, R.C. Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture Part I: Motivation and Philosophy. In Proceedings of the the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Amsterdam, The Netherlands, 12–15 March 2008.
6. Sparrow, R. Twenty Seconds to Comply: Autonomous Weapons Systems and the Recognition of Surrender. *Int'l L. Stud. Ser. US Naval War Col.* **2015**, *91*, 699.
7. Umbrello, S. Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data Cogn. Comput.* **2019**, *3*, 5. [CrossRef]
8. Van de Poel, I.; Kroes, P. Can technology embody values? In *The Moral Status of Technical Artefacts*; Springer: Dordrecht, The Netherlands, 2014; pp. 103–124.
9. Schmutz, M.; Borges, O.; Jesus, S.; Borchard, G.; Perale, G.; Zinn, M.; Sips, Ä.A.; Soeteman-Hernandez, L.G.; Wick, P.; Som, C. A methodological safe-by-design approach for the development of nanomedicines. *Front. Bioeng. Biotechnol.* **2020**, *8*, 258. [CrossRef] [PubMed]
10. Cawthorne, D.; Cenci, A. Value Sensitive Design of a Humanitarian Cargo Drone. In Proceedings of the 2019 International Conference on Unmanned Aircraft Systems, (ICUAS), Atlanta, GA, USA, 11–14 June 2019.
11. Davis, J.; Nathan, L.P. Value Sensitive Design: Applications, Adaptations, and Critiques. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Ed.; Springer: Dordrecht, The Netherlands, 2015.
12. Boscoe, B. Creating Transparency in Algorithmic Processes. *Delphi Interdiscip. Rev. Emerg. Technol.* **2019**, *2*, 12. [CrossRef]
13. Winkler, T.; Spiekermann, S. Twenty Years of Value Sensitive Design: A Review of Methodological Practices in Vsd Projects. *Ethics Inf. Technol.* **2018**, *21*, 1–5. [CrossRef]
14. Galliott, J.; Wyatt, A. Risks and Benefits of Autonomous Weapon Systems Perceptions among Future Australian Defence Force Officers. *J. Indo-Pac. Aff.* **2020**, *3*, 4.
15. Zhang, B.; Anderljung, M.; Kahn, L.; Dreksler, N.; Horowitz, M.C.; Dafoe, A. Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *arXiv* **2021**, arXiv:2105.05117. [CrossRef]
16. Galliott, J.; Cappuccio, M.; Wyatt, A. Taming the Killer Robot: Toward a Set of Ethical Principles for Military AI (Forthcoming). In press.
17. IPSOS. Six in Ten (61%) Respondents across 26 Countries Oppose the Use of Lethal Autonomous Weapons Systems, News Release. Available online: https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons (accessed on 22 January 2019).
18. Costa Rica Statement to 2020 Meeting of Group of Governmental Experts on Laws. In Proceedings of the Meeting of Governmental Experts of LAWS, Geneva, Switzerland, 21 September 2020; Available online: https://reachingcriticalwill.org/disarmament-fora/ccw/2020/laws/statements (accessed on 15 July 2021).
19. Holland, M.A. *The Black Box, Unlocked: Predictability and Understand-ability in Military AI.*; United Nations Institute for Disarmament Research: Geneve, Switzerland, 2020.
20. Jacobs, N.; Huldtgren, A. Why Value Sensitive Design Needs Ethical Commitments. In *Ethics and Information Technology*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 23–26.
21. Friedman, B.; Kahn, P.H.; Borning, A.; Huldtgren, A. Value Sensitive Design and Information Systems. In *The Ethics of Information Technologies*; Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M., Eds.; Springer: Dordrecht, The Netherlands, 2013.
22. Borning, A.; Muller, M. Next steps for value sensitive design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2012, Austin, TX, USA, 5–10 May 2012.
23. Convention on Certain Conventional Weapons. *Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System*; Group of Governmental Experts on LAWS: Geneva, Switzerland, 2019.
24. *Human Rights Watch: Stopping Killer Robots Country Positions on Banning Fully Autonomous Weapons and Retaining Human Controll*; Human Rights Watch: New York, NY, USA, 2021.
25. Brazil, Chile, and Mexico. Elements for a Future Normative Framework Conducive to a Legally Binding Instrument to Address the Ethical Humanitarian and Legal Concerns Posed by Emerging Technologies in the Area of (Lethal) Autonomous Weapons (Laws), News Release. Available online: https://documents.unoda.org/wp-content/uploads/2021/06/Brazil-Chile-Mexico.pdf (accessed on 15 July 2021).
26. Bolivarian Republic of Venezuela. "Reflections by the Bolivarian Republic of Venezuela on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (Laws) and the Mandate of the Group of Governmental Experts (Gge) Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to have Indiscriminate Effects". 2020. Available online: https://documents.unoda.org/wp-content/uploads/2020/09/20200901-Venezuela.pdf (accessed on 15 July 2021).

27. Republic of Philippines. "Commentary of the Republic of the Philippines on the Normative and Operational Framework in Emerging Technologies in the Area of Lethal Autonomous Weapon System. (Convention on Certain Conventional Weapons", 2021). Available online: https://documents.unoda.org/wp-content/uploads/2021/06/Philippines-.pdf (accessed on 15 July 2021).
28. Finland. Elements for Possible Consensus Recommendations; LAWS Group of Governmental Experts of the High Contracting Parties to the Convention on Certain Conventional Weapons. 2021. Available online: https://documents.unoda.org/wp-content/uploads/2021/06/Finland.pdf (accessed on 15 July 2021).
29. Commonwealth of Australia. *Convention on Certain Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems*; United Nations: New York, NY, USA, 2021.
30. Samuel Moncada. Baku Declaration of the 18th Midterm Ministerial Meeting of the Non-Aligned Movement. In *General Assembly*; United Nations: Baku, Republic of Azerbaijan, 2018.
31. Human Rights Watch. Us: Trump Administration Abandons Landmine Ban, News Release. Available online: https://www.hrw.org/news/2020/01/31/us-trump-administration-abandons-landmine-ban (accessed on 31 January 2020).
32. Tilman, R. The Nuclear Weapons Ban Treaty Is Groundbreaking, Even If the Nuclear Powers Haven't Signed. *The Conversation*. 21 January 2021. Available online: https://theconversation.com/the-nuclear-weapons-ban-treaty-is-groundbreaking-even-if-the-nuclear-powers-havent-signed-153197 (accessed on 15 July 2021).
33. Kania, E.B. *Chinese Military Innovation in Artificial Intelligence, Testimony to the US-China Economic and Security Review Commission*; Center for a New American Security: Wasington, DC, USA, 2019.
34. Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapons Systems in the Context of the Purposes and Objectives of the Convention . In *Convention on Certain Conventional Weapons*; Group of Governmental Experts on Lethal Autonomous Weapons: Geneva, Switzerland, 2018.
35. Russian Federation. Potential Opportunities and Limitations of Military Uses of Lethal Autonomous Weapons Systems," United Nations, CCW/GGE.1/2019/WP/1, (Convention on Certain Conventional Weapons, 2019). Available online: https://reachingcriticalwill.org/disarmament-fora/ccw/2019/laws/documents (accessed on 15 July 2021).
36. Adamsky, D. *The Culture of Military Innovation*; Stanford University Press: Stanford, CA, USA, 2020.
37. Raska, M. *Military Innovation in Small States: Creating a Reverse Asymmetry*; Routledge: London, UK, 2015.
38. Horowitz, M.C. *The Diffusion of Military Power*; Princeton University Press: Princeton, NJ, USA, 2010.
39. Goldman, E.O.; Leslie, C.E. *The Diffusion of Military Technology and Ideas*; Stanford University Press: Stanford, CA, USA, 2003.