Chapter 8

Ethical Constraints and Contexts of Artificial Intelligent Systems in National Security, Intelligence, and Defense/Military Operations

John R. Shook, Tibor Solymosi and James Giordano

Abstract

Weapons systems and platforms guided by Artificial Intelligence can be designed for greater autonomous decision-making with less real-time human control. Their performance will depend upon independent assessments about the relative benefits, burdens, threats, and risks involved with possible action or inaction. An ethical dimension to autonomous Artificial Intelligence (aAI) is therefore inescapable. The actual performance of aAI can be morally evaluated, and the guiding heuristics to aAI decision-making could incorporate adherence to ethical norms. Who shall be rightly held responsible for what happens if and when aAI commits immoral or illegal actions? Faulting aAI after misdeeds occur is not the same as holding it morally responsible, but that does not mean that a measure of moral responsibility cannot be programmed. We propose that aAI include a "Cooperating System" for participating in the communal ethos within NSID/military organizations.

Keywords: Artificial intelligence; autonomous weapons; morality programming; ethical AI; communal ethos; ethics; military ethics

Introduction

Decision technologies and artificially intelligent (AI) systems are being considered for their potential utility in national security, intelligence, and defense (NSID) operations (Giordano, Kulkarni, & Farwell, 2014; Giordano & Wurzman, 2016; Hallaq, Somer, Osula, Ngo, & Mitchener-Nissen, 2017). Given this trend, it becomes increasingly important to recognize the

Artificial Intelligence and Global Security, 137–152
Copyright © 2020 Emerald Publishing Limited
All rights of reproduction in any form reserved
doi:10.1108/978-1-78973-811-720201008

capabilities, limitations, effects, and need for guidance and governance of different types of AI that can – and will – be employed in NSID settings. Perhaps one of the most provocative, if not contentious, issues is the development and use of autonomous AI (aAI) to direct targeting and engagement of weapons systems (Galliott, 2015). To be sure, accurate identification, selection, and engagement of targets involve acquisition, discrimination, and parsing of multifactorial information. Given that the end-goal of engagement is reduction or elimination of the targeted threat, it is critical to address and assess mechanisms and processes of signal discrimination and selection in terms of the technical effectiveness and the rectitude of action(s). If accurate target identification and threat-reduction are the defined ends (and regarded "goods") of these systems' operation, then decisional accuracy (i.e., identification-discrimination-action; or more simply stated: "acquire-aim-fire") represents an intrinsic aspect of any such functions (Arkin, 2010; Canning, 2006; Lin, Abney, & Jenkins, 2017). For an aAI system, the decision to engage or not engage axiomatically obtains and entails the necessity to independently parse information about relative benefit, burden, threat, and risk domains that are contingent upon potential action or inaction (Arkin, 2010; Lin, Abney, & Bekey, 2007). Thus, missional effectiveness can be seen as involving both technical and ethical dimension. By pondering if and how an aAI system could be ethical (Arkin, 2009; Asaro, 2006; Wallach, Allen, & Franklin, 2011), or asking why ethical aAI is needed (Canning, 2006; Danielson, 2011; Sharkey, 2011), the fundamental worry is about who shall be rightly held responsible for what happens if and when aAI errs and commits acts that are regarded as "wrong".

Posing these issues and asking such questions of ethics are essential to understand – and establish – standards of acceptable behavior in the contexts in which humans live. Ever more, such discussions must acknowledge the growing role(s) of technology. The hoped-for promise of AI lies in humanity's recognition of these systems as extensions and enhancements of human capability and activity, but rarely, if ever, as a complete replacement for human beings and their involvement in decisions that establish vital contingencies. Yet, the trajectory toward specific forms of aAI delegates at least some of this decisional and actional responsibility. Therefore, we accordingly ask, to what extent may aAI systems meaningfully participate in the moral evaluation of their actions?

Apropos the explicit point and purpose of ethics, any AI system cannot be properly evaluated beyond the contexts in which they are employed by humans – who uphold communal values and standards (i.e., morals and mores) of acceptable conduct. Therefore, the same holds true for the use of AI in NSID contexts. Senior NSID personnel are held responsible for the deployment of all instruments of intelligence and combat. Commanders want force to be applied intelligently, and AI will become an integral component of that paradigm, not a poor substitute for it.

The use of AI in NSID will entail discrimination and execution of high-impact decisions and actions at faster-than-human speeds, but that does not mean that military or legal responsibility vanishes from sight. The computational capabilities of AI in the field would be rendered useless by continual intervention by slow human thinking. In this light, AI systems of future battlescapes will necessarily be largely autonomous (Scharre, 2018). However, a caveat is warranted: using the term "autonomous" to connote "dangerously uncontrolled" or "dumbly robotic" is ignorant at best, and disingenuous at worst.

Delineating Major Types of AI

AI will have an expanding role in automating an increasing number of operations through efficient information-processing and decision-routinization capabilities. Asaro (2006, 2008) has described four levels of machine system agency, which proceed from systems "with moral significance" that can execute basic decisions that can be of moral importance; through systems with increasing moral intelligence in their use of extant ethical codes, to an apex system that obtains dynamic moral intelligence. This final, most advanced iteration would engage some type of Bayesian-like decisional processes to advance from an initially programmed set of ethical precepts, to one of its own formulation, as based upon its own interactions with situations and environments (Rao, 2011). Building upon the work of Asaro, we posit that five core types of AI systems can be distinguished, with a view toward NSID applications, with higher-functioning types building on the capabilities of lower-functioning types. These types of AI are:

(1) **Automation AI** – those systems incorporated within stationed computing platforms for information management.
(2) **Animate AI** – systems acting to assess, traverse, and modify their environs (by projection, motility, and/or mobility) under human operational control.
(3) **Autonomous AI** – systems conducting assigned activities without the need for real-time human operation.
(4) **Agentic AI** – systems selectively engaging their environment in autonomously flexible pursuit of assigned outcomes.
(5) **Autopoetic AI** – systems that adapt decision heuristics for improved planning and execution of agential conduct.

A "human-supervised AI system" (of any kind 1–5) allows human operators to inaugurate, periodically direct, and terminate engagements under all conditions, including system failure. Of note is that those systems that temporarily act with a degree of independence do not constitute an entirely independent system. To be sure, types 3, 4, and 5 are autonomous, in increasingly complex ways, and terminology differs somewhat across AI industries and national governments (UNIDIR, 2017; USDoD, 2012). In the context of NSID use, we shall refer to an AI system that is able to decide upon and execute actions on its own – without direct and immediate operational control by a human – as an autonomous system. In contrast, an AI system that is only able to generate action when directly operated in real time by a human-in-the-loop is not autonomous.

Autonomous activity is certainly compatible with a degree of supervision. Humans are autonomous agents who receive guidance and direction, and are able to follow lawful commands.

Full independence and emancipation from humanity could arise with higher forms of intelligence, such as sapient or sentient AI. A sapient AI system autopoetically controls how it accomplishes goals and interacts with other intelligences, including interpreting or ignoring humans. A sentient AI would sufficiently sapient to comprehend its capabilities and construct its goals without needing or consulting humans (Giordano, 2015). Entirely unguided and unsupervised actors, whether biological or mechanical, could become renegade, and as such, be of dubious value to a nation, organization, or NSID operation.

Although critical works of technological foresight (Moravec, 1999; Wallach & Allen, 2009), and a corpus of science fiction depicting sapient or sentient AI pose profound ethical questions generated by sentient AI[1], it is highly unlikely that any NSID operation will consider the use of such systems in the proximate future.

Thus, acting with intelligent autonomy cannot – and arguably should not – be equated with behaving in an entirely independent way. Although the literal translation of the Greek *auto-nomos* is "self-rule," an individual who controls his/her own decisions and actions can still follow the social norms that one endorses. For example, an autonomous AI vehicle will follow driving rules and traffic laws, just as a good human driver does. Neither machine nor human should establish the habit of obeying only rules that they themselves create. Social norms, and moral norms in particular, promote cooperative practices. Habitual conformity with social norms, especially norms for communal welfare, are compatible with exercising individual autonomy within a defined group (of either "moral friends" or "moral strangers"; Engelhardt, 1996). Renegade deviance is only a crude kind of liberty. The conforming guidance of moral norms should contribute to productively autonomous existence.

Moral Guidance for Autonomous AI

Our practical question here is: To what extent could moral guidance be a component of ongoing supervision of an aAI (type 3–5) system, such as a ground robot, a flying drone, or a mechanical swarm? AI capable of degrading or destroying hostile forces shall be the primary subject in the following sections, unless otherwise specified. Raising the matter of moral guidance for AI is not just asking the question, "How should an ethics review of AI activity be imposed after military operations?" An external ethics review is usually too late to prevent or mitigate moral transgressions. Nor are we inquiring about preemptory ethical condemnations against aAI and their tactical performances in the battlefield (see here, Maas, 2019). Neither preemptory nor post operational examinations may be sufficiently attuned to the particular ways that human-AI systems will harmoniously work to achieve tactical goals during NSID (and more specifically military) operations. Rather, if it is to be effective in practice, moral guidance be constituent to, and occur within human-AI relationships.

Premising that moral guidance cannot essentially involve AI, but only be about AI, makes an assumption that renders an AI system functionally amoral. On this assumption, moral humans must be responsible for amoral AI. The pairing of an ethically autonomous human with an intelligently autonomous system at most ensures that the actions of the AI system can be compared against particular standards of right conduct. The system performs its AI actions in the field while understanding nothing about norms; the human judges the deeds of the AI system according to human norms while performing nothing in the field. That the AI system does not participate in its moral evaluation is akin to being convicted in a legal trial *in absentia*.

This sort of external ethics review does not characterize the moral evaluation of NSID personnel by their commanding officers and high-ranking leadership. This is especially true in the military. All military personnel are instructed in, inculcated with, and instilled with respect for military moral virtues and rules (Lucas, 2016). Compliance may vary to some extent given the variance of human action in certain circumstances, but none could claim ignorance of duty, and everyone can understand how and why their conduct is evaluated according to military moral standards. Military ethics is primarily about internal (i.e., organizational) oversight, supervision, and compliance. All military personnel: (1) adopt the required norms of proper behavior in uniform; (2) expect other service members to satisfy norms while conducting group activities; and (3) cooperate with military evaluations of personnel conduct by those standards. These three conditions establish the foundation of the "military ethos."

In general, a community inculcates and sustains its internal moral ethos where all members: (1) prioritize meeting the behavioral norms of the group; (2) participate in activities so as to promote the fulfillment of those norms in the conduct of participants, and (3) provide cooperation with communal efforts to uphold those standards of conduct. A communal ethos will weaken and decay when some members prioritize their own selfish goals, allow others to degrade group achievement, and disregard attempts to maintain collective standards. A communal ethos is fortified when members exemplify conformity, expect cohesiveness, and enforce compliance.

Internal moral guidance flourishes within the framework of a communal ethos. No single member of the organization takes the responsibility for morality, except in the inclusive sense that they all do. Morality infuses the whole; it does not reside in any part taken separately. Trying to ask, "Which particular component of the group takes the moral responsibility?" will fail to discern how moral responsibility is distributed throughout the relationships and practices of constituent members. This "collective ethics" model of internal moral guidance is in strong contrast with the model of external moral criticism.

Locating Moral Responsibility

Ascertaining moral responsibility can be accomplished externally. External moral criticism assumes a standpoint outside a group to inspect its components, workings, and activities. Adding aAI to a group of (autonomous) human agents allows the question, "If the AI acts in a way that is morally wrong, where would moral responsibility rest?" Taking the components of that human-AI system separately, the question actually becomes: "Shall moral responsibility reside with the human side, or the AI side?"

Emphasizing how any AI is merely an amoral tool is a short cut for placing the full burden of responsibility upon the human side. Claims that "AI in itself is neither good nor bad, but is only used rightly or wrongly by human users" establish that humans are fully responsible for aAI. Canning (2006) has argued that a constant construct across any spectrum of system autonomy should be the principle that machine systems must have either the involvement of a human operator or explicit human authorization to act against (i.e., neutralize) a human target. To paraphrase Canning (2006), machines can target other machines, while only humans may target humans. Yet full human responsibility is not a simple matter to confirm.

Moral responsibility for an outcome in a situation implies having some degree of control in that situation. If moral responsibility for an outcome resides with the party having the most control over that situation, where does moral responsibility reside in a human-AI system? A human *in*-the-loop who is continuously monitoring and orienting the AI system has significant control over its activities, and hence considerable moral responsibility for its activity, including any destructive or lethal actions. In contrast, while a human "*on*-the-loop" is periodically monitoring how an aAI system is performing, its specific acts (including lethal acts) could be performed without explicit human direction.

Let us suppose that an aAI system does something that no human authorizes, suggests, or even envisions. Clearly, there is no human oversight and control of that independent action. If absence of control implies lack of moral responsibility, no human supervisor is morally responsible for an aAI system's independent action. The aAI controlled its action, not a human; so, we could infer the aAI must be held morally responsible here. Yet, we opine that the assignment of moral responsibility to an aAI may be premature. Even sophisticated AI systems may not really "know" or "care" about right and wrong, and chastisement and punishment of AI seem pointless. Assigning moral blame could be more corrective than retributive, but fixing an AI system's programming does not first require assigning *moral* responsibility.

A paradox looms. We expect greater autonomy to be linked with greater responsibility. However, as Arkin (2010) and Bataoel (2011) have noted, increasing system autonomy does not inherently provide or confer parameters for the development and/or execution of moral decisions and ethical actions.

For example, while the aAI system is operating fairly independently to fulfill tactical goals, a human "on-the-loop" may bear little to no moral responsibility, and the AI system never has any moral responsibility. Conjoining human autonomy and AI autonomy in a human-AI system appears to only diminish or eliminate moral responsibility altogether. Ethics should be able to discern where moral responsibility resides and propose how to improve the capacity for moral responsibility. Where human-AI systems are concerned, ethics from an external standpoint would need to work harder at isolating and crediting full moral responsibility.

Applying ethical theories to aAI system cannot proceed from a presumption that the system is already an individual agent bearing moral responsibility. Presumptive responsibility must be attached to human agents. The "ethics of AI" is typically conceived simply as the "ethics of using AI." Consequentialist and deontological approaches offer accordingly distinctive analyses. Employing human-independent AI can be supported by consequentialist views, such as utilitarian arguments weighing the benefits of off-loading work that is too demanding for humans against the possible harms that could be (or are) incurred. However, the liberating humanity from labors could bring new forms of degradation or enslavement. Although sapient AI rebelling against its human masters is still the stuff of science fiction, deontological concerns can be raised about human dependency upon aAI's dominion of safety and security. Whether victories are gained or lost, the prospect of perpetual and escalating AI-driven and AI-executed warfare will challenge thresholds and tolerances for military conflict, and precepts of human values, freedoms, and rights (Asaro, 2008; Howlader & Giordano, 2013).

Parameters and boundaries for deploying aAI could be roughly determined by adjusting balances and compromises among utilitarian and deontological factors (Howlader & Giordano, 2013). But that is not the ethical issue pursued in this essay. Regardless of what idealized bounds are imaginable, actual AI will be constructed and deployed under real-world circumstances. If ethics proves unable to discern moral responsibility within in the human-AI system, where that abundance of autonomy appears to eliminate moral responsibility, there would be little point to conducting a moral inquiry and investigation into wrongful conduct on the part of any individual internal to the system. All the same, external ethical scrutiny can be applied to the system as a whole. How should AI be used by humans?

If the key issue for deontological and/or consequentialist address is solely the ethics of *using* aAI, there would be an initial dichotomization dividing humanity (the responsible moral agents) from aAI (mechanisms barely envisionable as agents). By separating AI, the design of these systems' programming becomes paramount, wherein "moral" prescriptions, proscriptions, and constraints could potentially be encoded (Franklin, 1995). Furthermore, so long as deontological and consequentialist approaches remain focused upon what the singular agent should and should not do, ethics is led toward developing some sort of "algorithm" or "application" for morality, which could run in real-time to direct and restrain the actions of an aAI system (see here, Casebeer, 2003, 2017). Thus, if we take the "artificial" in "artificial intelligence" to infer that AI is imitating human intelligence, then perhaps individual moral intelligence could be modeled and installed.

Programming AI Morality

The notion of programming an AI system to be good is almost as old as the idea of robotics itself. However, some serious obstacles block the road – or offer the proverbial dilemma of the "choice of the path taken" to designing moral programming. This is because deontological and consequentialist approaches to morality can point in differing directions. The means to utilitarian ends may fail to be right or just; and duty and righteousness for their own sake might not increase the overall good. Substituting other ethical theories cannot avoid this "ethical divergence" problem. Reliance on just one ethical theory can linearize moral thinking, but nonsubscribers to that theory need not agree that the thought process and answers it brings are right. It is impossible to say how AI could be as ethical as humans, so long as humans disagree about how to be ethical. This issue undergirds proposed aAI development in China, as the general idea is that the ambiguity of human moral and ethical decisional processes could (and should) be resolved through the integration of aAI system that conforms to centralized tenets of control, keeping the human element on-the-loop, but not necessarily in-the-loop. As Kania has noted (2018), the Chinese military is required to "remain a staunch force for upholding the Chinese Communist Party's ruling position," and would not tolerate any AI system that behaves in ways contrary to this tenet.

Autonomous AI can surely be "programmed" for moral conduct – indeed, many such moral programs are imaginable and at minimum, feasible. Moral proliferation is the real obstacle, and competing ethical theories are part of that problem, not a resolution. There is no optimal design solution: any compromise among ethical approaches would be practically indistinguishable from immorality, from the standpoint of one ethical theory or several theories. Any and all morally programmed AI would likely be categorized as evil by one subset of humanity or another. This is a familiar human situation: one righteous group (by its own ethical definition) is perceived as malevolent by another (according to that group's ethical standards). To this point, we are fond of paraphrasing MacIntyre's (1988) query: what good; whose justice; which rationality? It is irrelevant whether moral relativism has (or lacks) the endorsement of philosophical ethics; the world where humanity lives is a scene of genuine moral and ethical disagreement that cannot be ignored.

As long as humans disagree with each other, and with "moral" AI, about what is truly ethical conduct, there is no way to generally determine how any AI could be as ethical as humanity. The suboptimal alternative is to concede that any "moral AI" could only be as moral as the chosen ethical approach taken by a specific subgroup of humanity. That is, for a given human community with its own ethos, there could be, in principle, locally optimal moral programming for a AI (*vide supra*).

The implementation of local "morality programming" for AI is one thing; creating morally responsible AI is quite another. Even if high confidence could be bestowed upon a suitably complex programming for moral AI behavior, does that mean that this AI can now bear some moral responsibility for its actions? If we still feel uncomfortable crediting AI with any moral responsibility, we are forced to look elsewhere, specifically, to the human side of the human-AI system. Yet the humans on-the-loop seem even less responsible for whatever happens while the AI pursues its mission and makes its decisions with a high degree of independence. That tactical and moral independence lends the appearance of (perhaps even greater) autonomy to the AI side, yet moral responsibility remains as elusive as ever. No matter how we survey the human-AI system, the AI component appears only to have the status of an amoral tool.

Perhaps then, the humans responsible for the moral programming should be held responsible for any immoral behavior enacted by the programmed AI system. However, those programmers are even more distant, in a causal and control sense, from the concrete actions that the AI takes in the field. As well, training and testing moral programming within hypothetical scenarios can rarely, if ever, duplicate the unanticipatable contingencies of real-world engagements (e.g., against inherently unpredictable other AI systems and human agents). When an AI system performs oddly, in what seems to be an erratic or almost chaotic manner, such unwanted behavior is considered to be "accidental"; yet occasional accidents are almost inevitable due to the high complexity of aAI systems' structures and functions.

Simply put, that AI accidents will happen is no accident. The dynamically recursive and looped networks and integrated systems of aAI platforms, so tightly coupled high speed computation and communication, will produce nonlinear, cascading, and concatenating decisions and acts occurring too fast for human comprehension or intercession. The growing adoption of "deep learning" and auto-reprogramming will only make AI more inscrutable. If AI were slow enough for humans "on-the-loop" to always understand what AI is doing and why AI is doing it, then that AI would be practically worthless as an asset for gaining tactical advantage in the field. That advantage is amplified by expecting AI systems to make ever-faster tactical decisions with incomplete and inconclusive data, further increasing the odds of unpredictable and erratic behaviors (Scharre, 2016). Such eccentricities are acceptable features under certain NSID conditions. Yet, in any case, there is no good reason to hold human programmers morally responsible for all aAI decisions and actions.

To summarize matters so far, programming aAI systems for morality and moral responsibility confronts serious obstacles. Viewed externally, little moral responsibility can be discerned on either the human side or the AI side of a human-AI system, especially as AI functions more autonomously. Viewed internally, such obstacles are greatly reduced, but contextual dependence is increased. Morality programming for aAI can be designed for functioning within the context of a community's ethos.

Ethics in Context and Community

An ethics of and for a human-AI system, like any study of human-AI relationships, should apply network and systems principles (Liu, 2019). Making AI ethical cannot be simply about programming the AI for morality in imitation of human morality in a general sense. For human beings, morality is a complicated arena replete with tough dilemmas as well as simplistic platitudes. The amount of abstract rationality applied to moral thinking is not the issue. Just as humans are not atomistic rational beings outside of their bodies and their environments (inclusive of other biological organisms encountered and tools available), AI is and always will be similarly situated. This situatedness calls for an ethics appropriate to what AI is about within the human context. Ethics is a tool, and is grounded to the enterprise in which it shall be used (for overview of this construct in NSID contexts, see Tennison, Giordano, & Moreno, 2017). Regardless of whether a human or a machine is employing a deontological algorithm or a calculation for utility, such thinking is unhelpful if an actual problematic situation does not call for that tool.

Tools are extensions of human capacities. Ethics provides systems of rules and rationalizations to guide, govern, and in many cases justify moral decisions and actions. AI is also a tool – to extend human intelligence. Extending the reach of intelligence is not just a matter of duplication, or of adding one more intelligence to the world. Like ethics, the extensive capacity and power of AI is to enable decisions and actions that affect humans in interaction on and across a variety of scales. Thus, it functions as *social intelligence*. An autonomous system or machine is often presented atomistically, as a complete unit with hard and distinct boundaries between itself and the world. Irrespective of where it is in the world, it remains the same unit, the same machine, the same agency. This conception of intelligence and agency as being *discretely* autonomous is a myth, not merely about AI, but about humanity as well (Wurzman & Giordano, 2009).

The atomistic conception of autonomy is a pre-Darwinian view that takes the intelligence – whether called *psyche*, *nous*, soul, mind, consciousness, or simply "the self" is beside the point – as fully formed and final. In this view, moral reasoning is undertaken by a lone agent (regardless of who or what this agent happens to be) in order to render an "objective" moral judgment. After Darwin, the notion of such an idealized abstract intelligence was no longer tenable. Human intelligence is a bricolage of cognitive abilities, able to be more or less adaptive to contingent practical demands (Anderson, 2014; Johnson, 2014). Like human nature, human intelligence is not fixed and final. Rather, it is historical, situated in and embedded with multiscalar physical environments that are dynamic and changing in response to and with human activity (Wurzman & Giordano, 2009). This dynamic transaction is well coordinated by cybernetic functions of the nervous system: activity is coordinated and governed via feedback processes that modify feedforward anticipations of the human bodily system.

The human body, of course, lives within a larger systems-of-systems that include other enbrained and encultured humans (and other organisms), which are embedded in a wider ecology of the natural environment (Flanagan, 1996, 2017; Giordano, Benedikter, & Kohls, 2012; Solymosi, 2014). Intelligence is extensionally and functionally relational, making use of and operating across all of these levels of systemic complexity. Locating responsible intelligence at solely one level or in a special node of the whole network is to commit a category mistake (viz. a form of erroneous mereological conceptualiztion; Bennett & Hacker, 2003; Giordano, Rossi, & Benedikter, 2013). Intelligence of every kind is thoroughly social. Responsible moral intelligence, whether human or AI, is therefore social in nature and communal in its exercise (Giordano, Becker, & Shook, 2016). If moral performance is evaluated apart from mission performance, then the communal ethos is being ignored and overridden, degrading the responsibility of everyone involved.

We began our inquiry of AI ethics by focusing on the question of to what extent AI systems for NSID service may meaningfully participate in the moral evaluation of their actions. We propose that the answer now becomes clear: For a given human community with its own ethos – such as any NSID organization – there could be, in principle, locally optimal moral programming for aAI. To the extent that AI systems are tactically cooperating in the responsibly intelligent conduct of their mission activities, they already meaningfully participate in the moral conduct of the human-AI team.

Toward Synthesis: An AI "Cooperating System" for Ethics

Let us recall how a communal ethos (basically) functions within NSID/military organizations. Personnel: (1) adopt the required norms of proper behavior in uniform; (2) expect other service members to satisfy norms while conducting group activities; and (3) cooperate with community-focal evaluations of personnel conduct by those standards. This moral framework can be extended as a quasi-Gigerenzer (or other Bayesian) heuristic model to encompass aAI (co) operating within a human-AI team (Gigerenzer, 2000; Gigerenzer & Todd, 1999; for other heuristic models, see Brooks, 2002; Gams, Bohanec, & Cestnik, 1994; Hall, 2007; Roscheisen, Hofman, & Tresp, 1994). Autonomous AI will exemplify these three capacities (1, 2, 3) in a manner appropriate for AI, under all conditions:

Moral AI (1). AI will conform to the required norms of cooperative behavior while in service. Programming for heuristics of collaborative teamwork will function at the level of a "cooperating system" that hierarchically rests on the core operating system, preventing any noncooperative decision to reach the mechanical stage of actual action. All other programming is designed for functioning within that "cooperating system," and no other programming has the ability to engage any mechanical operation of the AI (e.g., no other programming by itself can engage an on-board weapon or use the AI system itself as a weapon).

Moral AI (2). The AI Cooperating System (AICS) will ensure that all on-board programming satisfies the communal norms of appropriate tactical conduct, inclusive of moral conduct. The AICS is empowered to temporarily halt or terminate any other programming functioning within its virtual environment, if persistently noncooperative (and possibly immoral) decisions are generated by it. Human approval for this AICS noncooperation override will *not* be required, although human notification should be promptly provided.

Moral AI (3). The AICS, whether operating as onboard computing or cloud-based computing (NB: ideally both, in case of communication disruptions), will provide real-time tactical data relaying its (own) cooperative performance (including moral conduct) for continual or periodic human review. The AICS will be reprogrammable for improvement and retraining of its heuristics for evaluation of AI cooperativeness. Importantly, the AICS will not be self-reprogrammable, but can only be modified/upgraded by designated NSID personnel.

Taken together, Moral AI (1–3) would serve as responsible entities designed for upholding the communal ethos of the NSID/military organization served. Because moral guidance will be embedded into human-AI teamwork, aAI and its AICS would exhibit and exemplify moral responsibility. The entirety of the human-AI system can achieve conformity with the NSID/military ethics of its communal ethos, since moral guidance is distributed throughout the human-AI system. Distributed intelligence, if it is responsible intelligence, will fulfill communal morality. In this context, AI ethics will be NSID/military ethics, conforming to the NSID/military ethos. An AI in NSID/military service will participate in its capacity to meet moral standards of conduct during NSID/military engagements.

Conclusion

In this chapter, we have investigated some minimal conditions and criteria for designing autonomous and responsible AI systems able to exemplify key NSID/military virtues. Virtuous responsibility is compatible with autonomous activity. Autonomy is not simply freedom from responsibility. Individual autonomy need not be antithetical or contrary to group responsibility. This communal ethos model of virtuous responsibility is exemplified in teamwork under good leadership. NSID/military organizations are prominent examples of communal ethos, wherein aAI systems can perform commendable service.

A core NSID/military virtue is loyalty, which requires trusting relationships. It is likely that humans will have understandable difficulty wholly trusting aAI, at least at first. But loyalty is a "two-way street," which is a colloquial way of describing trust-in-loyalty as being about a group's character, not just an individual characteristic. Thus, just as cooperative intelligence is distributed, responsible intelligence must also be extended and distributed. From the perspective of humans observing the conduct of AI, fulfilling that cooperative responsibility will allow AI to display key NSID/military virtues. In this way, we believe that NSID/military AI can be ethical AI. Of course, given ongoing international efforts in AI research, development, and use, as before (Giordano, 2013; Lanzilao, Shook, Benedikter, & Giordano, 2013; Tennison et al., 2017) we must once again – and perhaps consistently – ask: which military; what ethic? Such inquiry is important to define the possibilities and problems generated by uses of increasingly aAI, and establish process(es) to inform realistic and relevant perspectives, guidelines, and policies for direction and governance (Danielson, 2011). And we opine that apace with developments in AI, this endeavor should be – and remain – an interdisciplinary, international work-in-progress.

Acknowledgments

Dr. James Giordano's work is supported in part by the Henry M. Jackson Foundation for Military Medicine; Leadership Initiatives; and federal funds UL1TR001409 from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Clinical and Translational Science Awards Program (CTSA), a trademark of the Department of Health and Human Services, part of the Roadmap Initiative, "Re-Engineering the Clinical Research Enterprise." The views presented in this chapter do not necessarily reflect those of the US Department of Defense, US Special Operations Command, the Defense Advanced Research Projects Agency, or the authors' supporting institutions and organizations.

References

Anderson, M. L. (2014). *After phrenology: Neural re-use and the interactive brain*. Cambridge, MA: MIT Press.

Arkin, R. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine*, *28*(1), 30–33.

Arkin, R. (2010). The case for ethical autonomy in unmanned system. *Journal of Military Ethics*, *9*(4), 332–341.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics*, *6*(12), 9–16.

Asaro, P. M. (2008). How just could a robot war be? In A. Briggle, K. Waelbers, & P. A. E. Brey (Eds.), *Current issues in computing and philosophy*. Amsterdam: IOS Press.

Bataoel, V. (2011). On the use of drones in military operations in Libya: Ethical, legal and social issues. *Synesis: A Journal of Science, Technology, Ethics and Policy*, *2*(1), 69–76.

Bennett, M. R., & Hacker, P. M. S. (2003). *The philosophical foundations of neuroscience*. Oxford: Blackwell.

Brooks, R. (2002). *Robot: The future of flesh and machines*. London: Penguin.

Canning, J. S. (2006, September). A concept of operations for armed autonomous systems. Presented at the third annual disruptive technology conference, Washington, DC.

Casebeer, W. D. (2003). *Natural ethical facts. Evolution, connectionism and moral cognition*. Cambridge, MA: Bradford/MIT Press.

Casebeer, W. D. (2017, June). The case for an ethical machine system. Lecture presented at the Neuroethics Network Meeting, Paris, France.

Danielson, P. (2011). Engaging the public in the ethics of robots for war and peace. *Philosophy and Technology*, *24*(3), 239–249.

Engelhardt, H. T. (1996). *The foundations of bioethics* (2nd ed.). New York, NY: Oxford University Press.

Flanagan, O. (1996). Ethics naturalized: Ethics as human ecology. In L. May, M. Friedman, & A. Clark (Eds.), *Mind and morals: Essays on ethics and cognitive science*. Cambridge, MA: MIT Press.

Flanagan, O. (2017). *The geography of morals: Varieties of moral possibility*. New York, NY: Oxford University Press.

Franklin, S. (1995). *Artificial minds*. Cambridge, MA: MIT Press.

Galliott, J. (2015). *Military robots: Mapping the moral landscape*. New York, NY: Routledge.

Gams, M., Bohanec, M., & Cestnik, B. (1994). A schema for using multiple knowledge. In S. J. Hanson, T. Petsche, M. Kearns, & R. L. Rivest (Eds.), *Computational learning theory and natural learning systems*. Cambridge, MA: MIT Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Giordano, J. (2013). Respice finem: Historicity, heuristics and guidance of science and technology on the 21st century world stage. *Synesis: A Journal of Science, Technology, Ethics and Policy*, *4*, 1–4.

Giordano, J. (2015). Conscious machines? Trajectories, possibilities, and neuroethical considerations. *Artificial Intelligence Journal*, *5*(1), 11–17.

Giordano, J., Becker, K., & Shook, J. R. (2016). On the "neuroscience of ethics" – Approaching the neuroethical literature as a rational discourse on putative neural processes of moral cognition and behavior. *Journal of Neurology and Neuromedicine*, *1*(6), 32–36.

Giordano, J., Benedikter, R., & Kohls, N. B. (2012). Neuroscience and the importance of a neurobioethics: A reflection upon Fritz Jahr. In A. Muzur & H. M. Sass (Eds.), *Fritz Jahr and the foundations of integrative bioethics*. Münster; Berlin: LIT Verlag.

Giordano, J., Kulkarni, A., & Farwell, J. (2014). Deliver us from evil? The temptation, realities, and neuroethico-legal issues of employing assessment neurotechnologies in public safety initiatives. *Theoretical Medicine and Bioethics*, *35*(1), 73–89.

Giordano, J., Rossi, P. J., & Benedikter, R. (2013). Addressing the quantitative and qualitative: A view to complementarity – From the synaptic to the social. *Open Journal of Philosophy*, *3*(4), 1–5.

Giordano, J., & Wurzman, R. (2016). Integrative computational and neurocognitive science and technology for intelligence operations: Horizons of potential viability, value and opportunity. *STEPS: Science, Technology, Engineering and Policy Studies*, *2*(1), 34–38.

Hall, J. S. (2007). *Beyond AI*. Amherst, NY: Prometheus Books.

Hallaq, B., Somer, T., Osula, A. M., Ngo, T., & Mitchener-Nissen, T. (2017). Artificial intelligence within the military domain and cyber warfare. In *Proceedings of 16th European conference on cyber warfare and security*. Dublin: Academic Conferences and Publishing International Limited.

Howlader, D., & Giordano, J. (2013). Advanced robotics: Changing the nature of war and thresholds and tolerance for conflict – Implications for research and policy. *The Journal of Philosophy, Science and Law*, *13*, 1–19.

Johnson, M. (2014). *Morality for humans: Ethical understanding from the perspective of cognitive science*. Chicago, IL: University of Chicago Press.

Kania, E. (2018, April 17). China's strategic ambiguity and shifting approach to lethal autonomous weapons systems. Lawfare. Retrieved from https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems. Accessed on September 29, 2019.

Lanzilao, E., Shook, J., Benedikter, R., & Giordano, J. (2013). Advancing neuroscience on the 21st century world stage: The need for and a proposed structure of an internationally relevant neuroethics. *Ethics in Biology, Engineering and Medicine*, *4*(3), 211–229.

Lin, P., Abney, K., & Bekey, G. A. (2007). *Robot ethics. The ethical and social implications of robotics*. Cambridge, MA: MIT Press.

Lin, P., Abney, K., & Jenkins, R. (Eds.). (2017). *Robot ethics 2.0: From autonomous cars to artificial intelligence*. New York, NY: Oxford University Press.

Liu, H. Y. (2019). From the autonomy framework towards networks and systems approaches for 'autonomous' weapons systems. *Journal of International Humanitarian Legal Studies*, *10*(10), 1163.

Lucas, G. (2016). *Military ethics – What everyone needs to know*. Oxford: Oxford University Press.

Maas, M. (2019). Innovation-proof global governance for military artificial intelligence. *Journal of International Humanitarian Legal Studies*, *10*(10), 129–157.

MacIntyre, A. (1988). *Whose justice? Which rationality?* Notre Dame, IN: University of Notre Dame Press.

Moravec, H. (1999). *Robot: Mere machine to transcendent mind*. New York, NY: Oxford University Press.

Rao, R. P. N. (2011). Neural models of Bayesian belief propagation. In K. Doya (Ed.), *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: The MIT Press.

Roscheisen, M., Hofman, R., & Tresp, V. (1994). Incorporating prior knowledge into networks of locally-tuned units. In S. J. Hanson, T. Petsche, M. Kearns, & R. L. Rivest (Eds.), *Computational learning theory and natural learning systems*. Cambridge, MA: MIT Press.

Scharre, P. (2016). *Autonomous weapons and operational risk*. Washington, DC: Center for a New American Security. Retrieved from https://s3.amazonaws.com/files.cnas.org/. Accessed on September 30, 2019.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. New York, NY: W. W. Norton & Company.

Sharkey, N. (2011). Automating warfare: Lessons learned from the drones. *Journal of Law, Information and Science*, *21*(2), 140–154.

Solymosi, T. (2014). Moral first aid for a neuroscientific age. In T. Solymosi & J. R. Shook (Eds.), *Neuroscience, neurophilosophy, and pragmatism: Brains at work with the world*. New York, NY: Palgrave Macmillan.

Tennison, M., Giordano, J., & Moreno, J. (2017). Security threat versus aggregated truths: Ethical issues in the use of neuroscience and neurotechnology for national security. In J. Illes & S. Hossein (Eds.), *Neuroethics: Anticipating the future*. Oxford: Oxford University Press.

UNIDIR (United Nations' Institute for Disarmaments Research). (2017). *Weaponization of increasingly autonomous technology: Concerns, characteristics and definitional approaches* (No. 6). New York, NY: UNIDR Resources. Retrieved from https://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf. Accessed on September 1, 2019.

USDoD (US Department of Defense). (2012). DoD Directive 3000.09: Autonomy in weapons systems. Retrieved from https://www.hsdl.org/?view&did=726163. Accessed on September 1, 2019.

Wallach, W., & Allen, C. (2009). *Moral machines*. Oxford: Oxford University Press.

Wallach, W., Allen, C., & Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness*, *30*(1), 177–192.

Wurzman, R., & Giordano, J. (2009). Explanation, explanandum, causality and complexity: A consideration of mind, matter, neuroscience, and physics. *NeuroQuantology*, *7*(3), 368–381.