Autonomous Military Robotics: Risk, Ethics, and Design

Prepared for: US Department of Navy, Office of Naval Research

Prepared by: Patrick Lin, Ph.D.
George Bekey, Ph.D.
Keith Abney, M.A.

Ethics + Emerging Sciences Group at
California Polytechnic State University, San Luis Obispo

Prepared on: December 20, 2008

Version: 1.0.9

This work is sponsored by the Department of the Navy, Office of Naval Research, under awards # N00014-07-1-1152 and N00014-08-1-1209.

Table of Contents

Preface

This report is designed as a preliminary investigation into the risk and ethics issues related to *autonomous military systems*, with a particular focus on battlefield robotics as perhaps the most controversial area. It is intended to help inform policymakers, military personnel, scientists, as well as the broader public who collectively influence such developments. Our goal is to raise the issues that need to be consider in responsibly introducing advanced technologies into the battlefield and, eventually, into society. With history as a guide, we know that foresight is critical to both mitigate undesirable effects as well as to best promote or leverage the benefits of technology.

In this report, we will present: the presumptive case for the use of autonomous military robotics; the need to address risk and ethics in the field; the current and predicted state of military robotics; programming approaches as well as relevant ethical theories and considerations (including the Laws of War, Rules of Engagement); a framework for technology risk assessment; ethical and social issues, both near- and far-term; and recommendations for future work.

This work is sponsored by the US Department of the Navy, Office of Naval Research, under Awards # N00014-07-1-1152 and N00014-08-1-1209, whom we thank for its support and interest in this important investigation. We also thank California Polytechnic State University (Cal Poly, San Luis Obispo) for its support, particularly the College of Liberal Arts and the College of Engineering.

We are indebted to Colin Allen (Indiana Univ.), Peter Asaro (Rutgers Univ.), and Wendell Wallach (Yale) for their counsel and contributions, as well as to a number of colleagues—Ron Arkin (Georgia Tech), John Canning (Naval Surface Warfare Center), Ken Goldberg (IEEE Robotics and Automation Society; UC Berkeley), Patrick Hew (Defence Science and Technology Organization, Australia), George R. Lucas, Jr. (US Naval Academy), Frank Chongwoo Park (IEEE Robotics and Automation Society; Seoul National Univ.), Lt. Col. Gary Sargent (US Army Special Forces; Cal Poly), Noel Sharkey (Univ. of Sheffield, UK), Rob Sparrow (Monash Univ., Australia), and others—for their helpful discussions. We also thank the organizations mentioned herein for use of their respective images. Finally, we thank our families and nation's military for their service and sacrifice.

Patrick Lin
Keith Abney
George Bekey

December, 2008

1. Introduction

> "No catalogue of horrors ever kept men from war. Before the war you always think that it's not you that dies. But you will die, brother, if you go to it long enough."—Ernest Hemingway [1935, p.156]

Imagine the face of warfare with autonomous robotics: Instead of our soldiers returning home in flag-draped caskets to heartbroken families, autonomous robots—mobile machines that can make decisions, such as to fire upon a target, without human intervention—can replace the human soldier in an increasing range of dangerous missions: from tunneling through dark caves in search of terrorists, to securing urban streets rife with sniper fire, to patrolling the skies and waterways where there is little cover from attacks, to clearing roads and seas of improvised explosive devices (IEDs), to surveying damage from biochemical weapons, to guarding borders and buildings, to controlling potentially-hostile crowds, and even as the infantry frontlines.

These robots would be 'smart' enough to make decisions that only humans now can; and as conflicts increase in tempo and require much quicker information processing and responses, robots have a distinct advantage over the limited and fallible cognitive capabilities that we *Homo sapiens* have. Not only would robots expand the battlespace over difficult, larger areas of terrain, but they also represent a significant force-multiplier—each effectively doing the work of many human soldiers, while immune to sleep deprivation, fatigue, low morale, perceptual and communication challenges in the 'fog of war', and other performance-hindering conditions.

But the presumptive case for deploying robots on the battlefield is more than about saving human lives or superior efficiency and effectiveness, though saving lives and clearheaded action during frenetic conflicts are significant issues. Robots, further, would be unaffected by the emotions, adrenaline, and stress that cause soldiers to overreact or deliberately overstep the Rules of Engagement and commit atrocities, that is to say, war crimes. We would no longer read (as many) news reports about our own soldiers brutalizing enemy combatants or foreign civilians to avenge the deaths of their brothers in arms—unlawful actions that carry a significant political cost. Indeed, robots may act as objective, unblinking observers on the battlefield, reporting any unethical behavior back to command; their mere presence as such would discourage all-too-human atrocities in the first place.

Technology, however, is a double-edge sword with both benefits and risks, critics and advocates; and autonomous military robotics is no exception, no matter how compelling the case may be to pursue such research. The worries include: where responsibility would fall in cases of unintended or unlawful harm, which could range from the manufacturer to the field commander to even the machine itself; the possibility of serious malfunction and robots gone wild; capturing and hacking of military robots that are then unleashed against us; lowering the threshold for entering conflicts and wars, since fewer US military lives would then be at stake; the effect of such robots on squad cohesion, e.g., if robots recorded and reported back the soldier's every action; refusing an otherwise-legitimate order; and other possible harms.

We will evaluate these and other concerns within our report; and the remainder of this section will discuss the driving forces in autonomous military robotics and the need for 'robot ethics', as well as provide an overview of the report. Before that discussion, we should make a few introductory notes and definitions as follow.

1.1 Opening Remarks

First, in this investigation, we are *not* concerned with the question of whether it is even technically possible to make a perfectly-ethical robot, i.e., one that makes the 'right' decision in every case or even most cases. Following Arkin, we agree that an ethically-infallible machine ought not to be the goal now (if it is even possible); rather, our goal should be more practical and immediate: to design a machine that *performs better than* humans do on the battlefield, particularly with respect to reducing unlawful behavior or war crimes [Arkin, 2007]. Considering the number of incidences of unlawful behavior—and by 'unlawful' we mean a violation of the various Laws of War (LOW) or Rules of Engagement (ROE), which we also will discuss later in more detail—this appears to be a low standard to satisfy, though a profoundly important hurdle to clear. To that end, scientists and engineers need not first solve the daunting task of creating a truly 'ethical' robot, at least in the foreseeable future; rather, it seems that they only need to program a robot to act in compliance with the LOW and ROE (though this may not be as straightforward and simply as it first appears) or act ethically in the specific situations in which the robot is to be deployed.

Second, we should note that the purpose of this report is not to encumber research on autonomous military robotics, but rather to help responsibly guide it. That there should be two faces to technology—benefits and risk—is not surprising, as history shows, and is not by itself an argument against that technology.[1] But ignoring those risks, or at least only reactively addressing them and

[1] Biotechnology, for instance, promises to reduce world hunger by promoting greater and more nutritious agricultural and livestock yield; yet continuing concerns about the possible dissemination of bio-engineered seeds (or 'Frankenfoods') into the wild, displacing native plants and crops, have prompted the industry to move more cautiously [e.g., Thompson, 2007]. Even Internet technologies, as valuable as they have been in connecting us to

waiting for public reaction, seems to be unwise, given that it can lead (and, in the case of biotech foods, has led) to a backlash that stalls forward progress.

That said, it is surprising to note that one of the most comprehensive and recent reports on military robotics, *Unmanned Systems Roadmap 2007-2032*, does not mention the word 'ethics' once nor risks raised by robotics, with the exception of one sentence that merely acknowledges that "privacy issues [have been] raised in some quarters" without even discussing said issues [US Department of Defense, 2007, p. 48]. While this omission may be understandable from a public relations standpoint, again it seems short-sighted given lessons in technology ethics, especially from our recent past. Our report, then, is designed to address that gap, proactively and objectively engaging policymakers and the public to head off a potential backlash that serves no one's interests.

Third, while this report focuses on issues related to autonomous military *robotics*, the discussion may apply equally well and overlap with issues related to autonomous military *systems*, i.e., computer networks. Further, we are focusing on *battlefield* or lethal applications, as opposed to robotics in manufacturing or medicine even if they are supported by military programs (such as the Battlefield Extraction Assist Robot, or BEAR, that carries injured soldiers from combat zones), for several reasons as follow. The most contentious military robots will be the weaponized ones: "Weaponized unmanned systems is a highly controversial issue that will require a patient 'crawl-walk-run' approach as each application's reliability and performance is proved" [US Department of Defense, 2007, p. 54]. Their deployment is inherently about human life and death, both intended and unintended, so they immediately raise serious concerns related to ethics (e.g., does just-war theory or the LOW/ROE allow for deployment of autonomous fighting systems in the first place?) as well as risk (e.g., malfunctions and emergent, unexpected behavior) that demand greater attention than other robotics applications.

Also, though a relatively small number of military personnel is ever exposed on the battlefield, loss of life and property during armed conflict has non-trivial political costs, never mind environmental and economic costs, especially if 'collateral' or unintended damage is inflicted and even more so if it results from abusive, unlawful behavior by our own soldiers. How we prosecute a war or conflict receives particular scrutiny from the media and public, whose opinions influence military and foreign policy even if those opinions are disproportionately drawn from events on the battlefield, rather than on the many more developments outside the military theater. Therefore, though autonomous battlefield or weaponized robots may be years away and account for only one segment of the entire military robotics population, there is much practical value in sorting through their associative issues sooner rather than later.

---

information, social networks, etc., and in making new ways of life possible, reveal a darker world of online scams, privacy violations, piracy, viruses, and other ills; yet no one suggests that we should do away with cyberspace [e.g., Weckert, 2007].

Fourth and finally, while our investigation here is supported by the US Department of the Navy, Office of Naval Research, it may apply equally well to other branches of military service, all of which are also developing robotics for their respective needs. The range of robotics deployed or under consideration by the Navy, however, is exceptionally broad, with airborne, sea surface, underwater, and ground applications.[2] Thus, it is particularly fitting for the Department of the Navy to support one of the first dedicated investigations on the risk and ethical issues arising from the use of autonomous military robotics.

1.2 Definitions

To the extent that there are no standard, universally-accepted definitions of some of the key terms we employ in this report, we will need to stipulate those working definitions here, since it is important that we ensure we have the same basic understanding of those terms at the outset. And so that we do not become mired in debating precise definitions here, we provide a detailed discussion or justification for our definitions in 'Appendix A: Definitions'.

Robot (particularly in a military context). *A powered machine that (1) senses, (2) thinks (in a deliberative, non-mechanical sense), and (3) acts.*

Most robots are and will be mobile, such as vehicles, but this is not an essential feature; however, some degree of mobility is required, e.g., a fixed sentry robot with swiveling turrets or a stationary industrial robot with movable arms. Most do not and will not carry human operators, but this too is not an essential feature; the distinction becomes even more blurred as robotic features are integrated with the body. Robots can be operated semi- or fully-autonomously but cannot depend entirely on human control: for instance, tele-operated drones such as the Air Force's Predator unmanned aerial vehicle would qualify as robots to the extent that they make some decisions on their own, such as navigation, but a child's toy car tethered to a remote control is not a robot since its control depends entirely on the operator. Robots can be expendable or recoverable, and can carry a lethal or non-lethal payload. And robots can be considered as agents, i.e., they have the capacity to act in a world, and some even may be moral agents, as discussed in the next definition.

Autonomy (in machines). *The capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.*

[2] The only applications not covered by the Department of the Navy appear to be underground- and space-based, including sub-orbital missions, which may understandably fall outside their purview.

This is to say that, we are herein not interested in issues traditionally linked to autonomy that require a more robust and precise definition, such as the assignment of political rights and moral responsibility (as different from legal responsibility) or even more technical issues related to free will, determinism, personhood, and whether machines can even 'think'—as important as those issues are in philosophy, law, and ethics. But in the interest of simplicity, we will stipulate this definition, which seems acceptable in a discussion limited to human-created machines. This term also helps elucidate the second criterion of 'thinking' in our working definition of a robot. Autonomy is also related to the concept of moral agency, i.e., the ability to make moral judgments and choose one's actions accordingly.

Ethics (construed broadly for this report). *More than normative issues, i.e., questions about what we should or ought to do, but also general concerns related to social, political, and cultural impact as well as risk arising from the use of robotics.*

As a result, we will cover all these areas in our report, not just philosophical questions or ethical theory, with the goal of providing some relevant if not actionable insights at this preliminary stage. We will also discuss relevant ethical theories in more detail in section 3 (though this is not meant to be a comprehensive treatment of the subject).

1.3 Market Forces and Considerations

Several industry trends and recent developments—including high-profile failures of *semi*-autonomous systems, as perhaps a harbinger of challenges with more advanced systems—highlight the need for a technology risk assessment, as well as a broader study of other ethical and social issues related to the field. In the following, we will briefly discuss seven primary market forces that are driving the development of military robotics as well as the need for a guiding ethics; these roughly map to what have been called 'push' (technology) and 'pull' (social and cultural) factors [US Department of Defense, 2007, p.44].

1. *Compelling military utility*. US defense organizations are attracted to the use of robots for a range of benefits, some of which we have mentioned above. A primary reason is to replace us less-durable humans in "dull, dirty, and dangerous" jobs [US Department of Defense, 2007, p.19]. This includes: extended reconnaissance missions, which stretch the limits of human endurance to its breaking point; environmental sampling after a nuclear or biochemical attack, which had previously led to deaths and long-term effects on the surveying teams; and neutralizing IEDs, which have caused over 40% of US casualties in Iraq since 2003 [Iraq Coalition Casualty Count, 2008]. While official statistics are difficult to locate, news organizations report

that the US has deployed over 5,000 robots in Iraq and Afghanistan, which have neutralized 10,000 IEDs by 2007 [CBS, 2007].

Also mentioned above, military robots may be more discriminating, efficient, and effective. Their dispassionate and detached approach to their work could significantly reduce the instances of unethical behavior in wartime—abuses that negatively color the US prosecution of a conflict, no matter how just the initial reasons to enter the conflict are, and carry a high political cost.

2. *US Congressional deadlines*. Clearly, there is a tremendous advantage to employing robots on the battlefield, and the US government recognizes this. Two key Congressional mandates are driving the use of military robotics: by 2010, one-third of all operational deep-strike aircraft must be unmanned, and by 2015, one-third of all ground combat vehicles must be unmanned [National Defense Authorization Act, 2000]. Most, if not all, of the robotics in use and under development are semi-autonomous at best; and though the technology to (responsibly) create fully autonomous robots is near but not quite in hand, we would expect the US Department of Defense to adopt the same, sensible 'crawl-walk-run' approach as with weaponized systems, given the serious inherent risks.

   Nonetheless, these deadlines apply increasing pressure to develop and deploy robotics, including autonomous vehicles; yet a 'rush to market' increases the risk for inadequate design or programming. Worse, without a sustained and significant effort to build in ethical controls in autonomous systems, or even to discuss the relevant areas of ethics and risk, there is little hope that the early generations of such systems and robots will be adequate, making mistakes that may cost human lives. (This is related to the 'first-generation' problem we discuss in sections 6 and 7, that we won't know exactly what kind of errors and mistaken harms autonomous robots will commit until they have already done so.)

3. *Continuing unethical battlefield conduct*. Beyond popular news reports and images of purportedly unethical behavior by human soldiers, the US Army Surgeon General's Office had surveyed US troops in Iraq on issues in battlefield ethics and discovered worrisome results. From its summary of findings, among other statistics: "Less than half of Soldiers and Marines believed that non-combatants should be treated with respect and dignity and well over a third believed that torture should be allowed to save the life of a fellow team member. About 10% of Soldiers and Marines reported mistreating an Iraqi non-combatant when it wasn't necessary...Less than half of Soldiers and Marines would report a team member for unethical behavior...Although reporting ethical training, nearly a third of Soldiers and Marines reported encountering ethical situations in Iraq in which they didn't know how to respond" [US Army Surgeon General's Office, 2006]. The most recent survey by the same organization reported similar results [US Army Surgeon General's Office, 2008].

Wartime atrocities have occurred since the beginning of human history, so we are not operating under the illusion that they can be eliminated altogether (nor that armed conflicts can be eliminated either, at least in the foreseeable future). However, to the extent that military robots can considerably reduce unethical conduct on the battlefield—greatly reducing human and political costs—there is a compelling reason to pursue their development as well as to study their capacity to act ethically.

4. *Military robotics failures*. More than theoretical problems, military robotics have already failed on the battlefield, creating concerns with their deployment (and perhaps even more concern for more advanced, complicated systems) that ought to be addressed before speculation, incomplete information, and hype fill the gap in public dialogue.

   In April 2008, several TALON SWORDS units—mobile robots armed with machine guns—in Iraq were reported to be grounded for reasons not fully disclosed, though early reports claim the robots, without being commanded to, trained their guns on 'friendly' soldiers [e.g., Page, 2008]; and later reports denied this account but admitted there had been malfunctions during the development and testing phase prior to deployment [e.g., Sofge, 2008]. The full story does not appear to have yet emerged, but either way, the incident underscores the public's anxiety—and the military's sensitivity—with the use of robotics on the battlefield (also see 'Public perceptions' below).

   Further, it is not implausible to suggest that these robots may fail, because it has already happened elsewhere: in October 2007, a semi-autonomous robotic cannon deployed by the South African army malfunctioned, killing nine 'friendly' soldiers and wounding 14 others [e.g., Shachtman, 2007]. Communication failures and errors have been blamed for several unmanned aerial vehicle (UAV) crashes, from those owned by the Sri Lanka Air Force to the US Border Patrol [e.g., BBC, 2005; National Transportation Safety Board, 2007]. Computer-related technology in general is especially susceptible to malfunctions and 'bugs' given their complexity and even after many generations of a product cycle; thus, it is reasonable to expect similar challenges with robotics.

5. *Related civilian systems failures*. On a similar technology path as autonomous robots, civilian computer systems have failed and raised worries that can carry over to military applications. For instance, such civilian systems have been blamed for massive power outages: in early 2008, Florida suffered through massive blackouts across the entire state, as utility computer systems automatically shut off and rerouted power after just a small fire caused by a failed switch at one electrical substation [e.g., Padgett, 2008]; and in the summer 2003, a single fallen tree had triggered a tsunami of cascading computer-initiated blackouts that affected tens of millions of

customers for days and weeks across the eastern US and Canada, leaving practically no time for human intervention to fix what should have been a simple problem of stopping the disastrous chain reaction [e.g., US Department of Energy, 2004]. Thus, it is a concern that we also may not be able to halt some (potentially-fatal) chain of events caused by autonomous military systems that process information and can act at speeds incomprehensible to us, e.g., with high-speed unmanned aerial vehicles.

Further, civilian robotics are becoming more pervasive. Never mind seemingly-harmless entertainment robots, some major cities (e.g., Atlanta, London, Paris, Copenhagen) already boast driverless transportation systems, again creating potential worries and ethical dilemmas (e.g., bringing to life the famous thought-experiment in philosophy: should a fast-moving train divert itself to another track in order to kill only one innocent person, or continue forward to kill the five on its current path?). So there can be lessons for military robotics that can be transferred from civilian robotics and automated decision-making, and vice versa. Also, as robots become more pervasive in the public marketplace—they are already abundant in manufacturing and other industries—the broader public will become more aware of risk and ethical issues associated with such innovations, concerns that inevitably will carry over to the military's use.

6. *Complexity and unpredictability*. Perhaps robot ethics has not received the attention it needs, at least in the US, given a common misconception that robots will do only what we have programmed them to do. Unfortunately, such a belief is a sorely outdated, harking back to a time when computers were simpler and their programs could be written and understood by a single person. Now, programs with millions of lines of code are written by teams of programmers, none of whom knows the entire program; hence, no individual can predict the effect of a given command with absolute certainty, since portions of large programs may interact in unexpected, untested ways. (And even straightforward, simple rules such as Asimov's Laws of Robotics can create unexpected dilemmas [e.g., Asimov, 1950].) Furthermore, increasing complexity may lead to *emergent behaviors*, i.e., behaviors not programmed but arising out of sheer complexity [e.g., Kurzweil, 1999, 2005].

   Related major research efforts also are being devoted to enabling robots to learn from experience, raising the question of whether we can predict with reasonable certainty *what* the robot will learn. The answer seems to be negative, since if we could predict that, we would simply program the robot in the first place, instead of requiring learning. Learning may enable the robot to respond to novel situations, given the impracticality and impossibility of predicting all eventualities on the designer's part. Thus, unpredictability in the behavior of complex robots is a major source of worry, especially if robots are to operate in unstructured environments,

rather than the carefully-structured domain of a factory. (We will discuss machine learning further in sections 2 and 3.)

7. *Public perceptions*. From Asimov's science fiction novels to Hollywood movies such as *Wall-E*, *Iron Man, Transformers*, *Blade Runner, Star Wars, Terminator, Robocop, 2001: A Space Odyssey,* and *I, Robot* (to name only a few, from the iconic to recently released), robots have captured the global public's imagination for decades now. But in nearly every one of those works, the use of robots in society is in tension with ethics and even the survival of humankind. The public, then, is already sensitive to the risks posed by robots—whether or not those concerns are actually justified or plausible—to a degree unprecedented in science and technology. Now, technical advances in robotics are catching up to literary and theatrical accounts, so the seeds of worry that have long been planted in the public consciousness will grow into close scrutiny of the robotics industry with respect to those ethical issues, e.g., the book *Love and Sex with Robots* published late last year that reasonably anticipates human-robot relationships [Levy, 2007].

Given such investments, questions, events, and predictions, it is no wonder that more attention is being paid to robot ethics, particularly in Europe [e.g., Veruggio, 2007]. An entire conference dedicated to the issue of ethics in autonomous military systems—one of the first we have seen, if not the first of its kind—was held in late February 2008 in the UK [Royal United Services Institute (RUSI) for Defence and Security Studies, 2008], in which experts reiterated the possibility that robots might commit war crimes or be turned on us by terrorists and criminals [RUSI, 2008: Noel Sharkey and Rear Admiral Chris Parry's presentations, respectively; also, Sharkey, 2007a, and Asaro, 2008]. Robotics is a particularly thriving and advanced industry in Asia: South Korea is the first (and still only?) nation to be working on a 'Robot Ethics Charter' or a code of ethics to govern responsible robotics development and use, though the document has yet to materialize [BBC, 2007]. This summer, Taiwan played host to a conference about advanced robotics and its societal impacts [Institute of Electrical and Electronics Engineers (IEEE), 2008].

But the US is starting to catch up: some notable US experts are working on similar issues, which we will discuss throughout this report [Arkin, 2007; Wallach and Allen, 2008]. A January 2008 conference at Stanford University focused on technology in wartime, of which robot ethics was one notable session [Computer Professionals for Social Responsibility (CPSR), 2008]. In July 2008, the North American Computing and Philosophy (NA-CAP) conference at Indiana University focused a significant part of its program on robot ethics [NA-CAP, 2008]. Again, we intend for this report as an early, complementary step in filling the gap in robot-ethics research, both technical and theoretical.

1.4 Report Overview

Following this introduction, in section 2, we will provide a short background discussion on robotics in general and in defense applications specifically. We will survey briefly the current state of robotics in the military as well as developments in progress and anticipated. This includes several future scenarios in which the military may employ autonomous robots, which will help anchor and add depth to our discussions later on ethics and risk.

In section 3, we will discuss the possibility of programming in rules or a framework in robots to govern their actions (such as Asimov's Laws of Robotics). There are different programming approaches: top-down, bottom-up, and a hybrid approach [Wallach and Allen, 2008]. We also discuss the major (competing) ethical theories—deontology, consequentialism, and virtue ethics—that these approaches correspond with as well as their limitations.

In section 4, we consider an alternative, as well as a complementary approach, to programming a robot with an ethical behavior framework: to simply program it to obey the relevant Laws of War and Rules of Engagement. To that end, we also discuss the relevant LOW and ROE, including a discussion of just-war theory and related issues that may arise in the context of autonomous robots.

In section 5, continuing the discussion about law, we will also look at the issue of legal responsibility based on precedents related to product liability, negligence and other areas [Asaro, 2007]. This at least informs questions of risk in the near- and mid-term in which robots are essentially human-made tools and not moral agents of their own; but we also look at the case for treating robots as quasi-legal agents.

In section 6, we will broaden our discussion in providing a framework for technology risk assessment. This framework includes a discussion of the major factors in determining 'acceptable risk': consent, informed consent, affected population, seriousness, and probability [DesJardins, 2003].

In section 7, we will bring the various ethics and social issues discussed, and new ones, together in one location. We will survey a full range of possible risks and issues related to ethics, just-war theory, technical challenges, societal impact, and more. These contingencies and issues are important to have in mind in any complete assessment of technology risks.

Finally, in section 8, we will draw some preliminary conclusions, including recommendations for future, more detailed investigations. A bibliography is provided as section 9 of the report; and appendix A offers more detailed discussions on key definitions, as initiated in this section.

2. Military Robotics

The field of robotics has changed dramatically during the past 30 years. While the first programmable articulated arms for industrial automation were developed by George Devol and made into commercial products by Joseph Engleberger in the 1960s and 1970s, mobile robots with various degrees of autonomy did not receive much attention until the 1970s and 1980s. The first true mobile robots arguably were Elmer and Elsie, the electromechanical 'tortoises' made by W. Grey Walter, a physiologist, in 1950 [Walter, 1950]. These remarkable little wheeled machines had many of the features of contemporary robots: sensors (photocells for seeking light and bumpers for obstacle detection), a motor drive and built-in behaviors that enabled them to seek (or avoid) light, wander, avoid obstacles and recharge their batteries. Their architecture was basically reactive, in that a stimulus directly produced a response without any 'thinking.' That development first appeared in Shakey, a robot constructed at Stanford Research Laboratories in 1969 [Fikes and Nilsson, 1971]. In this machine, the sensors were not directly coupled to the drive motors but provided inputs to a 'thinking' layer known as the Stanford Research Institute Problem Solver (STRIPS), one of the earliest applications of artificial intelligence. The architecture was known as 'sense-plan-act' or 'sense-think-act' [Arkin, 1998].

Since those early developments, there have been major strides in mobile robots—made possible by new materials, faster, smaller and cheaper computers (Moore's law) and major advances in software. At present, robots move on land, in the water, in the air, and in space. Terrestrial mobility uses legs, treads, and wheels as well as snake-like locomotion and hopping. Flying robots make use of propellers, jet engines, and wings. Underwater robots may resemble submarines, fish, eels, or even lobsters. Some vehicles capable of moving in more than one medium or terrain have been built. Service robots, designed for such applications as vacuum cleaning, floor washing and lawn mowing, have been sold in large quantities in recent years. Humanoid robots, long considered only in science fiction novels, are now manufactured in various sizes and with various degrees of sophistication [Bekey, 2005]. Small toy humanoids, such as the WowWee Corporation's RoboSapien, have been sold in quantities of millions. More complex humanoids, such as the Honda ASIMO are able to perform numerous tasks. However, 'killer applications' for humanoid robots have not yet emerged.

There has also been great progress in the development of software for robots, including such applications as learning, interaction with humans, multiple robot cooperation, localization and navigation in noisy environments, and simulated emotions. We discuss some of these developments briefly in section 2.6 below.

During the past 20 years, military robotic vehicles have been built using all the modes of locomotion described above and making use of the new software paradigms [US Dept. Of Defense, 2007]. Military robots find major applications in surveillance, reconnaissance, location and destruction of mines and IEDs, as well as for offense or attack. The latter class of vehicles is equipped with weapons, which at the present time are fired by remote human controllers. In the following, we first summarize the state of the art in military robots, including both hardware and software, and then introduce some of the ethical issues which arise from their use. We concentrate on robots capable of lethal action—in that much of the concern with military robotics is tied to this lethality—and omit discussion of more innocuous machines such as the Army's Big Dog, a four legged robot capable of carrying several hundred pounds of cargo over irregular terrain. If at some future time such 'carry robots' are equipped with weapons, they may need to be considered from an ethical point of view.

2.1 Ground Robots

The US Army makes use of two major types of autonomous and semi-autonomous ground vehicles: large vehicles, such as tanks, trucks and HUMVEEs and small vehicles, which may be carried by a soldier in a backpack (such as the PackBot shown in Fig. 2.0a) and move on treads like small tanks [US Dept. Of Defense, 2007]. The PackBot is equipped with cameras and communication equipment and may include manipulators (arms); it is designed to find and detonate IEDs, thus saving lives (both civilian and military), as well as to perform reconnaissance. Its small size enables it to enter buildings, report on possible occupants, and trigger booby traps. Typical armed robot vehicles are (1) the Talon SWORDS (Special Weapons Observation Reconnaissance Detection System) made by Foster-Miller, which can be equipped with machine guns, grenade launchers, or anti-tank rocket launchers as well as cameras and other sensors (see Fig. 2.0b) and (2) the newer MAARS (Modular Advanced Armed Robotic System). While vehicles such as SWORDS and the newer MAARS are able to autonomously navigate toward specific targets through its global positioning system (GPS), at present the firing of any on-board weapons is done by a soldier located a safe distance away. Foster-Miller provides a universal control module for use by the warfighter with any of their robots. MAARS uses a more powerful machine gun than the original SWORDS. While the original SWORDS weighted about 150 lbs., MAARS weighs about 350 lbs. It is equipped with a new manipulator capable of lifting 100 lbs., thus enabling it to replace its weapon platform with an IED identification and neutralization unit.

Among the larger vehicles, the Army's Tank-Automotive Research, Development and Engineering Center (jointly with Foster-Miller) has developed the TAGS-CX, a 5,000-6,000 lb. amphibious vehicle. More recently, and jointly with Carnegie Mellon University, the Army has developed a 5.5 ton, six-

wheel unmanned vehicle known as the Crusher, capable of carrying 2,000 lbs. at about 30 mph and capable of withstanding a mine explosion; it is equipped with one or more guns (see figure 2.1).

(a) (b)

Fig. 2.0 Military ground vehicles: (a) PackBot (Courtesy of iRobot Corp.); (b) SWORDS (Courtesy of Foster-Miller Corp.)

Fig. 2.1 Military ground vehicle: The Crusher (Courtesy of US Army)

Both PackBot and Talon robots are being used extensively and successfully in Iraq and Afghanistan. Hence, we expect further announcements of UGV deployments in the near future. We are not aware of the use of armed sentry robots by the US military; however, they are used in South Korea

(developed by Samsung) and in Israel. The South Korean system is capable of interrogating suspects, identifying potential enemy intruders, and autonomous firing of its weapon.

DARPA supported two major national competitions leading to the development of autonomous ground vehicles. The 2005 Grand Challenge required autonomous vehicles to traverse portions of the Mojave desert in California. The vehicles were provided with GPS coordinates of way-points along the route, but otherwise the terrain to be traversed was completely unknown to the designers, and the vehicles moved autonomously at speed averaging 20 to 30 mph. In 2007, the Urban Challenge required autonomous vehicles to move in a simulated urban environment, in the presence of other vehicles and signal lights, while obeying traffic laws. While the winning automobiles from Stanford University and Carnegie Mellon University were not military in nature, the lessons learned will undoubtedly find their way into future generations of autonomous robotic vehicles developed by the Army and other services.

2.2 Aerial Robots

The US Army, Air Force, and Navy have developed a variety of robotic aircraft known as unmanned flying vehicles (UAVs).[3] Like the ground vehicles, these robots have dual applications: they can be used for reconnaissance without endangering human pilots, and they can carry missiles and other weapons. The services use hundreds of unarmed UAVs, some as small as a model airplane, to locate and identify enemy targets. An important function for unarmed UAVs is to serve as aerial targets for piloted aircraft, such as those manufactured by AeroMech Engineering in San Luis Obispo, CA, a company started by Cal Poly students. AeroMech has sold some 750 UAVs, ranging from 4 lb. battery-operated ones to 150 lb. vehicles with jet engines. Some reconnaissance UAVs, such as the Shadow, are launched by a catapult and can stay aloft all day. The best known armed UAVs are the semi-autonomous Predator Unmanned Combat Air Vehicles (UCAV) built by General Atomics (see Fig. 2.2a), which can be equipped with Hellfire missiles. Both the Predator and the larger Reaper hunter-killer aircraft are used extensively in Afghanistan. They can navigate autonomously toward targets specified by GPS coordinates, but a remote operator located in Nevada (or in Germany) makes the final decision to release the missiles. The Navy, jointly with Northrop Grumman, is developing an unmanned bomber with folding wings which can be launched from an aircraft carrier.

The military services are also developing very small aircraft, sometimes called Micro Air Vehicles (MAV) capable of carrying a camera and sending images back to their base. An example is the Micro

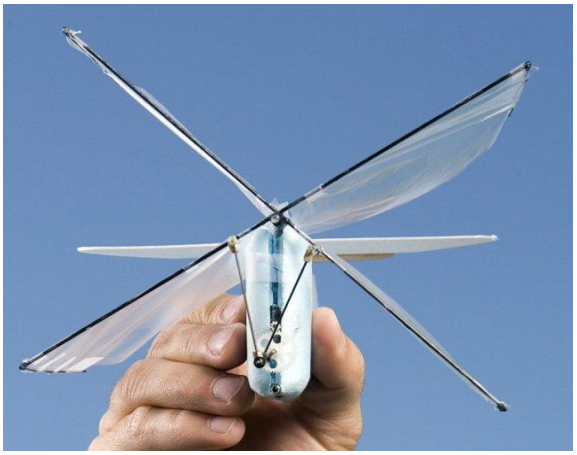
[3] Earlier versions of such vehicles were termed 'drones', which implied that they were completely under control of a pilot in a chaser aircraft. Current models are highly autonomous, receiving destination coordinates from only ground or satellite transmitters. Thus, because this report is focused on robots—machines that have some degree of autonomy—we do not use the term 'drone' here.

Autonomous Air Vehicle (MAAV; also called MUAV for Micro Unmanned Air Vehicle) developed by Intelligent Automation, Inc., which is not much larger than a human hand (see Fig. 2.2b).

(a) (b)

Fig. 2.2 Autonomous aircraft: (a) Predator (Courtesy of General Atomics Aeronautical Systems); (b) Micro unmanned flying vehicle (Courtesy of Intelligent Automation, Inc.)

Similarly, the University of Florida has developed an MAV with a 16-inch wingspan with foldable wings, which can be stored in an 8-inch x 4-inch container. Other AUVs include a ducted fan vehicle (see Fig. 2.3a) being used in Iraq, and vehicles with flapping wings, made by AeroVironment and others (Fig. 2.3b). While MAVs are used primarily for reconnaissance and are not equipped with lethal weapons, it is conceivable that the vehicle itself could be used in 'suicide' missions.



(a) (b)

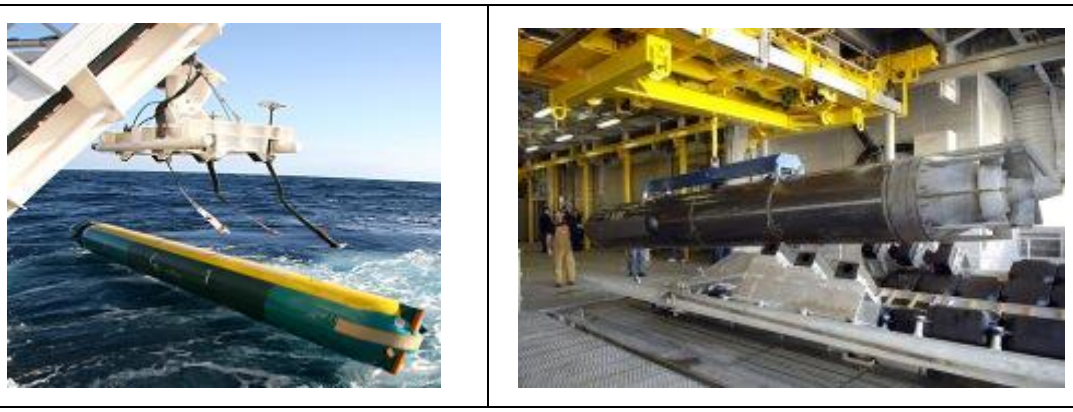
Fig. 2.3 Micro air vehicles: (a) ducted fan vehicle from Honeywell; (b) Ornithopter MAV with flapping wings made by students at Brigham Young University (Photo by Jaren Wilkey/BYU, used by permission)

Other flying robots either deployed or in development, including helicopters, tiny robots the size of a bumblebee, and solar-powered craft capable of remaining aloft for days or weeks at a time. Again, our objective here is not to provide a complete survey, but to indicate the wide range of mobile robots in use by the military services.

2.3 Marine Robots

Along with the other services, the US Navy has a major robotic program involving interaction between land, airborne, and seaborne vehicles [US Dept. of the Navy, 2004; US Dept. of Defense, 2007]. The latter include surface ships as well as Unmanned Underwater Vehicles (UUVs). Their applications include surveillance, reconnaissance, anti-submarine warfare, mine detection and clearing, oceanography, communications, and others. It should be noted that contemporary torpedoes may be classified as UUVs, since they possess some degree of autonomy.

As with robots in the other services, UUVs come in various sizes, from man-portable to very large. Fig. 2.4a shows Boeing's Long-term Mine Reconnaissance System (LMRS) which is dropped into the ocean from a telescoping torpedo launcher aboard the SV Ranger to begin its underwater surveillance test mission. LMRS uses two sonar systems, an advanced computer and its own inertial navigation system to survey the ocean floor for up to 60 hours. The LMRS shown in the figure is about 21 inches in diameter; it can be launched from a torpedo tube, operate autonomously, return to the submarine, and be guided into a torpedo-tube mounted robotic recovery arm. A large UUV, the Seahorse, is shown in Fig. 2.4b; this vehicle is advertised as being capable of 'independent operations', which may include the use of lethal weapons. The Seahorse is about 3 feet in diameter, 28 feet long, and weighs 10,500 lbs. The Navy plans to move toward deployment of large UUVs by 2010. These vehicles may be up to 3 to 5 feet in diameter, weighing perhaps 20,000 lbs.

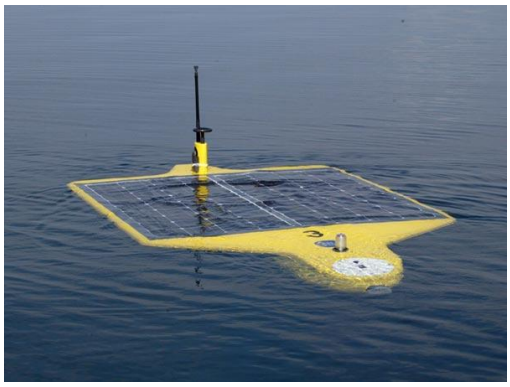


(a) (b)

Figure 2.4: (a) Long-term Mine Reconnaissance UUV (Courtesy of The Boeing Company); (b) Seahorse 3-foot diameter UUV (Courtesy of Penn State University)

Development of UUVs is not restricted to the US. Large UUV programs exist in Australia, Great Britain, Sweden, Italy, Russia, and other countries. Fig. 2.5a shows a UUV made in Great Britain by BAE Systems.

A solar-powered surface vehicle is shown in Fig. 2.5b. As with other military robots, most of the vehicles capable of delivering deadly force are currently human-controlled and not fully autonomous. However, the need for autonomy is great for underwater vehicles, since radio communication underwater is difficult. Many UUVs surface periodically to send and receive messages.



(a) (b)

Fig. 2.5: (a) Talisman UUV (Courtesy of BAE Systems);
(b) Solar powered surface vehicle (Courtesy of NOAA)

2.4 Space Robots

We believe that the US Armed Services have significant programs for the development of autonomous space vehicles: for advanced warning, defense against attacking missiles and possibly offensive action as well. However, there is very little information on these programs in publicly available sources. It is clear that the Air Force is building a major communication system in space, named Transformational Satellite Communication System (TSC). This system will interact with airborne as well as ground-based communication nodes to create a truly global information grid.

2.5 Immobile/Fixed Robots

To this point we have described a range of mobile robots used by the military: on earth, on and under the water, in the air, and in space. It should be noted that not all robots capable of lethal action are mobile; in fact, some are stationary, with only limited mobility (such as aiming of a gun). We consider a few examples of such robots in this section.

First, let us consider again why land mines and underwater mines, whether aimed at destruction of vehicles or attacks on humans (anti-personnel mines), are not properly robots. Whether buried in the ground or planted in the surf zone along the ocean shore, these systems are equipped with some sensing ability (since they can detect the presence of weight), and they 'act' by exploding. Their information processing ability is extremely limited, generally consisting only of a switch triggered by pressure from above. Given our definition of autonomous robots as consider in section 1 (as well as detailed in Appendix A), while such mines may be considered as autonomous, we do not classify them as robots since a simple trigger is not equivalent to the cognitive functions of a robot. If a landmine is considered a robot, one seems to be absurdly required to designate a trip wire as a robot too.

On the other hand, there are immobile or stationary weapons, both on land and on ships, which do merit the designation of robot, despite their lack of mobility (though they have some moving features, which satisfies our definition for what counts as a robot). An example of such a system is the Navy's Phalanx Close-In Weapon System (CIWS). CIWS is a rapid-fire 20mm gun system designed to protect ships at close range from missiles which have penetrated other defenses. The system is mounted on the deck of a ship; it is equipped with both search and tracking radars and the ability to rotate a turret in order to aim the guns. The information processing ability of the computer system associated with the radars is remarkable, since it automatically performs search, detecting, tracking, threat evaluation, firing, and kill-assessments of targets. Thus, the CIWS uses radar sensing of approaching missiles, identifies targets, tracks targets, makes the decision to fire, and then fires its guns, using solid tungsten bullets to penetrate the approaching target. The gun-and-radar turret can rotate in at least two degrees of freedom for target tracking, but the entire structure is immobile and fixed on the deck.

The US Army has also adopted a version of the Phalanx system to provide close-in protection for troops and facilities in Iraq, under the name 'Counter Rocket, Artillery, and Mortar' (C-RAM, or Counter-RAM). The system is mounted on the ground or, in some cases, on a train platform. The basic system operation is similar to that of the Navy system: it is designed to destroy incoming missiles at a relatively short range. However, since the system is located adjacent to or near civilian

facilities, there is major concern for collateral damage, e.g., debris or fragments of a disabled missile could land on civilians.

As a final example here, we cite the SGR-A1 sentry robot developed by Samsung Techwin Co. for use by the South Korean army in the Demilitarized Zone (DMZ) which separates North and South Korea. The system is stationary, designed to replace a manned sentry location. It is equipped with sophisticated color vision sensors that can identify a person entering the DMZ, even at night under only starlight illumination. Since any person entering the DMZ is automatically presumed to be an enemy, it is not necessary to separate friend from foe. The system is equipped with a machine gun, and the sensor-gun assembly is capable of rotating in two degrees of freedom as it tracks a target. The firing of the gun can be done manually by a soldier or by the robot in fully-automatic (autonomous) mode.

2.6 Robot Software Issues

In the preceding, we have presented the current state of some of the robotic hardware and systems being used and/or being developed by the military services. It is important to note that in parallel with the design and fabrication of new autonomous or semi-autonomous robotic systems, there is a great deal of work on fundamental theoretical and software implementation issues which also must be solved if fully autonomous systems are to become a reality [Bekey, 2005]. The current state of some of these issues is as follows:

2.6.1 Software Architecture

Most current systems use the so-called 'three level architecture', illustrated in Fig. 2.6. The lowest level is basically reflexive, and allows the robot to react almost instantly to a particular sensory input.

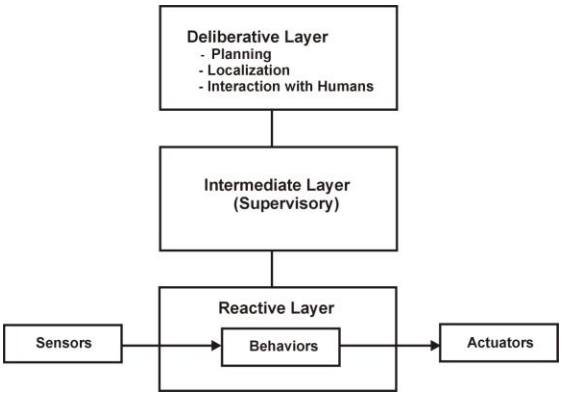


Figure 2.6. Typical three-level architecture for robot control

The highest level, sometimes called the Deliberative layer, includes Artificial Intelligence such as planning and learning, as well as interaction with humans, localization and navigation. The intermediate or 'supervisory' layer provides oversight of the reactive layer, and translates upper level commands as required for execution. Many recent developments have concentrated on increasing the sophistication of the 'deliberative' layer.

2.6.2 Simultaneous Localization and Mapping (SLAM)

An important problem for autonomous robots is to ascertain their location in the world and then to generate new maps as they move. A number of probabilistic approaches to this problem have been developed recently.

2.6.3 Learning

Particularly in complex situations it has become clear that robots cannot be programmed for all eventualities. This is particularly true in military scenarios. Hence, the robot must learn the proper responses to given stimuli, and its performance should improve with practice.

2.6.4 Multiple Robot System Architectures

Increasingly, it will become necessary to deploy multiple robots to accomplish dangerous and complex tasks. The proper architecture for control of such robot groups is still not known. For example, should they be organized hierarchically, along military lines, or should they operate in semi-autonomous sub-groups, or should the groups be totally decentralized?

2.6.5 Human-Robot Interaction

In the early days of robotics (and even today in certain industrial applications), robots are enclosed or segregated to ensure that they do not harm humans. However, in an increasing number of applications, humans and robots cooperate and perform tasks jointly. This is currently a major focus of research in the community, and there are several international conference devoted to Human-Robot Interaction (HRI).

2.6.6 Reconfigurable Systems

There is increasing interest (both for military and civilian applications) in developing robots capable of some form of 'shape-shifting.' Thus, in certain scenarios, a robot may be required to move like a

snake, while in others it may need legs to step over obstacles. Several labs are developing such systems.

2.7 Ethical Implications: A Preview

It is evident from the above survey that the Armed Forces of the United States are implementing the Congressional mandate described in section 1 of this report. However, as of this writing, none of the fielded systems has full autonomy in a wide context. Many are capable of autonomous navigation, localization, station keeping, reconnaissance and other activities, but rely on human supervision to fire weapons, launch missiles, or exert deadly force by other means; and even the Navy's CIWS operates in full-auto mode only as a reactive last line of defense against incoming missiles and does not proactively engage an enemy or target. Clearly, there are fundamental ethical implications in allowing full autonomy for these robots. Among the questions to be asked are:

- Will autonomous robots be able to follow established guidelines of the Laws of War and Rules of Engagement, as specified in the Geneva Conventions?
- Will robots know the difference between military and civilian personnel?
- Will they recognize a wounded soldier and refrain from shooting?

Technical answers to such questions are being addressed in a study for the US Army by professor Ronald Arkin from Georgia Institute of Technology—his preliminary report is entitled *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture* [Arkin 2007]—and other experts [e.g., Sharkey, 2008a]. In the following sections of our report, we seek to complement that work by exploring other (mostly non-technical) dimensions of such questions, specifically as they related to ethics and risk.

2.8 Future Scenarios

From the brief descriptions of the state of the art of robotics above, it is clear that the field is highly dynamic. Robotics is inherently interdisciplinary, drawing from advances in computer science, aerospace, electrical and mechanical engineering, as well as biology (to obtain models of sensing, processing and physical action in the animal kingdom), sociology, ergonomics (to provide a basis for the design and deployment of robot colonies), and psychology (to obtain a basis for human-robot interaction). Hence, discoveries in any of these fields will have an effect on the design of future robots and may raise new questions of risk and ethics. It would be useful, then, to anticipate possible future scenarios involving military robotics in order to more completely consider issues in risk and ethics, as follow:

2.8.1 Sentry/Immobile Robots

A future scenario may include robot sentries that guard not only military installations but also factories, government buildings, and the like. As these guards acquire increasing autonomy, they may not only challenge visitors ("Who goes there?") and ask them to provide identification but will be equipped with a variety of sensors for this purpose: vision systems, bar code readers, microphones, sound analyzers, and so on. Vision systems (and, if needed, fingerprint readers) along with large graphic memories may be used to perform the identification. More importantly, the guards will be equipped with weapons enabling them to arrest and, if necessary, to disable or kill a potential intruder who refuses to stop and be identified. Under what conditions will such lethal force be authorized? What if the robot confuses the identities of two people? These are only two of the many difficult ethical questions which will arise even in such a basically 'simple' task as guarding a gate and challenging visitors.

2.8.2 Ground Vehicles

We expect that future generations of Army ground vehicles, beyond the existing PackBots or SWORDS discussed in section 2.1 above, will feature significantly more and better sensors, better ordnance, more sophisticated computers, and associated software. Advanced software will be needed to accomplish several tasks, such as:

(a) Sensor fusion: More accurate situational awareness will require the technical ability to assign degrees of credibility to each sensor and then combine information obtained from them. For example, in the vicinity of a 'safe house', the robot will have to combine acoustic data (obtained from a variety of microphones and other sensors) with visual information, sensing of ground movement, temperature measurements to estimate the number of humans within the house, and so on. These estimates will then have to be combined with reconnaissance data (say from autonomous flying vehicles) to obtain a probabilistic estimate of the number of combatants within the house.

(b) Attack decisions: Sensor data will have to be processed by software that considers the applicable Rules of Engagement and Laws of War in order for a robot to make decisions related to lethal force. It is important to note that the decision to use lethal force will be based on probabilistic calculations, and absolute certainty will not be possible. If multiple robot vehicles are involved, the system will also be required to allocate functions to individual members of the group, or they will be required to negotiate with each other to determine their individual functions. Such negotiation is a current topic of much challenging research in robotics.

(c) Human supervision: We anticipate that autonomy will be granted to robot vehicles gradually, as confidence in their ability to perform their assigned tasks grows. Further, we expect to see learning algorithms that enable the robot to improve its performance during training missions. Even so, there will be fundamental ethical issues. For example, will a supervising warfighter be able to override a robot's decision to fire? If so, how much time will have to be allocated to allow such decisions? Will the robot have the ability to disobey a human supervisor's command, say in a situation where the robot makes the decision not to release a missile on the basis that its analysis leads to the conclusion that the number of civilians (say women and children) greatly exceeds the number of insurgents in the house?

2.8.3 Aerial Vehicles

Clearly, many of the same consideration that apply to ground vehicles will also apply to UFVs, with the additional complexity that arises from moving in three degrees of freedom, rather than two as on the surface of the earth. Hence, the UFV must sense the environment in the x, y, and z directions. The UFV may be required to bomb particular installations, in which case it will be governed by similar considerations to those described above. However, there may be others: for instance, an aircraft is generally a much more expensive system than a small ground vehicle such as the SWORDS. What evasive action should the vehicle undertake to protect itself? It should have the ability to return to base and land autonomously, but what should it do if challenged by friendly aircraft? Are there situations in which it may be justified in destroying friendly aircraft (and possibly killing human pilots) to ensure its own safe return to base? The UFV will be required to communicate with UGVs and to coordinate strategy when necessary. How should decisions be made if there is disagreement between airborne and ground vehicles? If there are hybrid missions that include both piloted and autonomous aircraft, who is in charge?

These are not a trivial question, since contemporary aircraft move at very high speeds, making the length of time required for decisions inadequate for human cognitive processes. In addition, vehicles may be of vastly different size, speed and capability. Further, under what conditions should a UFV be permitted to cross national boundaries in the pursuit of an enemy aircraft? Since national boundaries are not painted on the ground, the robot aircraft will have to rely on stored maps and GPS measurements, which may be faulty.

2.8.4 Marine Vehicles

Many of the same challenges that apply to airborne vehicles also apply to those traveling under water. Again, they must operate in multiple degrees of freedom. In addition, the sensory abilities of robot submarines will be quite different from those of ground or air vehicles, given the properties of water. Thus, sonar echoes can be used to identify the presence of underwater objects, but these

signals require interpretation. Assume that the robot submarine detects the presence of a surface vessel, which is presumed to carrying enemy weapons, as well as civilian passengers: under what conditions should the robot submarine launch torpedoes to destroy the surface vessel? It may be much more difficult to estimate the number of civilians aboard an iron ship than those present in a wooden house. How can the robot make intelligent decisions in the absence of critical information?

It is evident that the use of autonomous robots in warfare will pose a large number of ethical challenges. In the next sections, we discuss some programming approaches and their relationship to ethical theories, issues related to responsibility and law (including LOW/ROE), and expand on the various ethical and risk issues we have raised in the course of this report.

3. Programming Morality

What role might ethical theory play in defining the control architecture for semi-autonomous and autonomous robots used by the military? What moral standards or ethical subroutines should be implemented in a robot? This section explores the ways in which ethical theory may be helpful for implementing moral decision making faculties in robots.[4]

Engineers are very good at building systems to satisfy clear task specifications, but there is no clear task specification for general moral behavior, nor is there a single answer to the question of whose morality or what morality should be implemented in AI. However, military operations are conducted within a legal framework of international treaties as well as the nation's own military code. This suggests that the rules governing acceptable conduct of personnel might perhaps be adapted for robots; one might attempt to design a robot which has an explicit internal representation of the rules and strictly follows them.

A robotic code would, however, probably need to differ in some respects from that for a human soldier. For example, self-preservation may be less of a concern for the robotic system, both in the way it is valued by the military and in its programming. Furthermore, what counts as a strictly correct interpretation of the laws in a specific situation is itself likely to be a matter for dispute, and conflicts among duties or obligations will require assessment in light of more general moral principles. Regardless of what code of ethics, norms, values, laws, or principles are adopted for the design of an artificial moral agent (AMA), whether the system functions successfully will need to be evaluated through externally-determined criteria and testing.

3.1 From Operational to Functional Morality

Safety and reliability have always been a concern for engineers in their design of intelligent systems and for the military in its choice of equipment. Remotely-operated vehicles and semi-autonomous weapons systems used during military operations need to be reliable, and they should be destructive only when directed at designated targets. Not all robots utilized by the military will be deployed in combat situations, however, establishing as a priority that all intelligent systems are safe and do no harm to (friendly) military personnel, civilians, and other agents worthy of moral consideration.

[4] We thank and credit Wendell Wallach and Colin Allen for their contribution to many of the discussions here, drawn from their new book *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2008).

When robots with even limited autonomy must choose from among different courses of action, the concern for safety is transmuted into the need for the systems to have a capacity for making moral judgments. For robots that operate within a very limited context, the designers and engineers who build the systems may well be able to discern all the different options the robot will encounter and program the appropriate responses. The actions of such a robot are completely in the hands of the designers of the systems and those who choose to deploy them; these robots are *operationally moral*. They do not have, and presumably will not need, a capacity to explicitly evaluate the consequences of their actions. They will not need to evaluate which rules apply in a particular situation, nor need to prioritize conflicting rules.

However, three factors suggest that operational morality is not sufficient for many robotic applications: (1) the increasing autonomy of robotic systems; (2) the prospect that systems will encounter influences that their designers could not anticipate because of the complexity of the environments in which they are deployed, or because the systems are used in contexts for which they were not specifically designed; and (3) the complexity of technology and the inability of systems engineers to predict how the robots will behave under a new set of inputs.

The choices available to systems that possess a degree of autonomy in their activity and in the contexts within which they operate, and greater sensitivity to the moral factors impinging upon the course of actions available to them, will eventually outstrip the capacities of any simple control architecture. Sophisticated robots will require a kind of *functional morality*, such that the machines themselves have the capacity for assessing and responding to moral considerations. However, the engineers that design functionally moral robots confront many constraints due to the limits of present-day technology. Furthermore, any approach to building machines capable of making moral decisions will have to be assessed in light of the feasibility of implementing the theory as a computer program.

In the following, we will briefly examine several major theories—deontological (rule-based) ethics, consequentialism, natural law, social contract ethics, and virtue ethics—as possible ethical frameworks in robots. (A complete discussion of these theories and their relative plausibility is beyond the scope of this report and can be readily found in philosophical literature [e.g. University of San Diego, 2008].)

First, let us dismiss one important possibility: ethical relativism, or the position that there is no such thing as objectivity in ethical matters, i.e., what is right or wrong is not a matter of fact but a result of individual or cultural preferences. Even if it were true that ethics is relative to cultural preferences, this would have no bearing on a project to develop autonomous military robots, since the US military and its code of ethics would be the standard for our robots anyway, as opposed to programming

some other nation's morality into our machines. Further, we can expect that such robots will be employed only in specific environments, at least for the foreseeable future, which suggests a more limited, practical programming approach; so a broad or all-encompassing theory of ethics is not immediately urgent, and thus we need not settle the question of whether ethics is objective here.

That is, the idea of an autonomous general- or even multi-purpose robot (which might require a broad framework to govern a full range of possible actions) is much more distant than the possibility of an autonomous robot created for specific military-related tasks, such as patrolling borders or urban areas, or exercising lethal force in a carefully circumscribed battlefield. Given the limited operations of such robots, the initial ethical task will be sufficient to simply program in the suitable basic, relevant rules. In the next section, we will delineate the Laws of War and Rules of Engagement that would govern the robot's behavior; these laws already are established and codified, making programming easier (in theory). We will also offer challenges and further difficulties related to the approach of using the LOW and ROE as an ethical framework, and discuss longer-term issues that may arise as robots have greater autonomy and responsibility.

3.2 Overview: Top-Down and Bottom-Up Approaches

The challenge of building artificial moral agents (AMAs) might be understood as finding ways to implement abstract values within the control architecture of intelligent systems. Philosophers confronted with this problem are likely to suggest a top-down approach of encoding a particular ethical theory in software. This theoretical knowledge could then be used to rank options for moral acceptability. Psychologists confronted with the problem of constraining moral decision-making are likely to focus on the way a sense of morality develops in human children as they mature into adults. Their approach to the development of moral acumen is bottom-up in the sense that it is acquired over time through experience. The challenge for roboticists is to decide whether a top-down ethical theory or a bottom-up process of learning is the more effective approach for building artificial moral agents.

The study of ethics commonly focuses on top-down norms, standards, and theoretical approaches to moral judgment. From Socrates' dismantling of theories of justice to Kant's project of rooting morality within reason alone, ethical discourse has typically looked at the application of broad standards of morality to specific cases. According to these approaches, standards, norms, or principles are the basis for evaluating the morality of an action.

The term 'top-down' is used in a different sense by engineers, who approach challenges with a top-down analysis through which they decompose a task into simpler subtasks. Components are assembled into modules that individually implement these simpler subtasks, and then the modules

are hierarchically arranged to fulfill the goals specified by the original project.

In our discussion of machine morality, we use 'top-down' in a way that combines these two somewhat different senses from engineering and ethics. In our broader sense, a top-down approach to the design of AMAs is any approach that takes a specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory.

In the bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and is rewarded for behavior that is morally praiseworthy. In this manner, the artificial agent develops or learns through its experience. Unlike top-down ethical theories, which define what is and is not moral, ethical principles must be discovered or constructed in bottom-up approaches. Bottom-up approaches, if they use a prior theory at all, do so only as a way of specifying the task for the system, and not as a way of specifying an implementation method or control structure.

Engineers would find this top-down/bottom-up dichotomy to be rather simplistic given the complexity of many engineering tasks. However, the concepts of top-down and bottom-up task analysis are helpful in that they highlight two different roles for ethical theory in facilitating the design of AMAs.

3.3 Top-Down Approaches

Are ethical principles, theories, and frameworks useful in guiding the design of computational systems capable of acting with some degree of autonomy? Can top-down theories—such as utilitarianism, or Kant's categorical imperative, or even Asimov's laws for robots—be adapted practically by roboticists for building AMAs?

Top-down approaches to artificial morality are generally understood as having a set of rules that can be turned into an algorithm. These rules specify the duties of a moral agent or the need for the agent to calculate the consequences of the various courses of action it might select. The history of moral philosophy can be viewed as a long inquiry into the adequacy of any one ethical theory; thus, selecting any particular theoretical framework may not be adequate for ensuring an artificial agent will behave acceptably in all situations. However, one theory or another is often prominent in a particular domain, and for the foreseeable future most robots will function within limited domains of activity.

3.3.1 Top-Down Rules: Deontology

A basic grounding in ethical theory naturally begins with the idea that morality simply consists in following some finite set of rules: deontological ethics, or that morality is about simply doing one's duty. Deontological (duty-based) ethics presents ethics as a system of inflexible rules; obeying them makes one moral, breaking them makes one immoral. Ethical constraints are seen as a list of either forbidden or permissible forms of behavior. Kant's *Categorical Imperative (CI)* is typical of a deontological approach, as follows in its two main components:

CI(1) – This is often called the formula of universal law (FUL), which commands: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" [Kant, 1785, 4:421]. Alternatively, the CI also has been understood as that the relevant legislature should pass such a law mandating my action, i.e., a 'Universal Law of Nature.'

A maxim is a statement of one's intent or rationale: it is the answer to the query about why one did what was done. So Kant asserts that the only intentions that are moral are those that could be universally held; partiality has no place in moral thought. Kant also asserts that when we treat other people as a mere means to our ends, such action must be immoral; after all, we ourselves don't wish to be treated that way. Hence, when applying the CI in any social interaction, Kant provides a second formulation as a purported corollary:

CI(2) – Variously called the Humanity formulation of the CI, or the Means-Ends Principle, or the formula of the end in itself (FEI), it commands: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" [Kant, 1785, 4:429]. One could never universalize the treatment of another as a mere means to some other ends, claims Kant, in his explanation that CI(2) directly follows from CI(1). This formulation is credited with introducing the idea of 'respect' for persons; that is, respect for whatever it is that is essential to our Humanity, for whatever collective attributes are required for human dignity [Johnson, 2008].

A Kantian deontologist thus believes that acts such as stealing and lying are always immoral, because universalizing them creates a paradox. For instance, one cannot universalize lying without running into the 'Liar's paradox' (that it cannot be true that all statements are a lie); similarly, one cannot universalize stealing property without undermining the very concept of property. Kant's approach is widely influential but has problems of applicability and disregard for consequences.

3.3.2 Asimov's Laws of Robotics

Another deontological approach often comes to mind in investigating robot ethics: Asimov's Three

Laws of Robotics (he later added a fourth or 'Zeroth Law') are intuitively appealing in their simple demand to not harm or allow humans to be harmed, to obey humans, and to engage in self-preservation. Furthermore, the laws are prioritized to minimize conflicts. Thus, doing no harm to humans takes precedence over obeying a human, and obeying trumps self-preservation. However, in story after story, Asimov demonstrated that three simple hierarchically-arranged rules could lead to deadlocks when, for example, the robot received conflicting instructions from two people or when protecting one person might cause harm to others.

The original version of Asimov's Three Laws of Robotics are as follows: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey orders given to it by human beings, except where such orders would conflict with the First Law; (3); a robot must protect its own existence as long as such protection does not conflict with the First or Second Law [Asimov, 1950].

Asimov's fiction explored the implications and difficulties of the Three Laws of Robotics. It established that the first law was incomplete as stated, due to the problem of ignorance: a robot was fully capable of harming a human being as long as it did not know that its actions would result in (a risk of) harm, i.e., the harm was unintended. For example, a robot, in response to a request for water, could serve a human a glass of water teeming with bacterial contagion, or throw a human down a well, or drown a human in a lake, *ad infinitum*, as long as the robot was unaware of the risk of harm. One solution is to rewrite the first and subsequent laws with an explicit knowledge-qualifier: "A robot may do nothing that, to its knowledge, will harm a human being; nor, through inaction, knowingly allow a human being to come to harm" [Asimov, 1957]. But a clever criminal could divide a task among multiple robots, so that no one robot could even recognize that its actions would lead to harming a human, e.g., one robot places the dynamite, another attaches a length of cord to the dynamite, a third lights the cord, and so on. Of course, this simply illustrates the problem with deontological, top-down approaches, that one may follow the rules perfectly but still produce terrible consequences.

An additional difficulty is that the degree of risk makes a difference too, e.g., should robots keep humans from working near X-ray machines because of a small risk of cancer, and how would a robot decide? (Section 6 on risk assessment will explore this topic further). The 'through inaction' clause of Asimov's first law raises another issue: Wouldn't a robot have to constantly intervene to minimize all sorts of risks to humans, and never be able to perform its primary tasks? Asimov considers a modified First Law to solve this issue: (1´) A robot may not harm a human being. Removing the First Law's 'inaction' clause solves this problem, but t does so at the expense of creating an even greater one: a robot could initiate an action which would harm a human (for example, initiating an automatic firing sequence, then watching a noncombatant wander into the firing line) knowing that it was

capable of preventing the harm (by ceasing the automatic firing), but it may nevertheless fail to do so since it is now not strictly required to act.

Asimov later added a Zeroth Law [Asimov, 1985]—so named to continue the pattern of lower-numbered laws superseding in importance the higher-numbered laws—so that the Zeroth Law had highest priority and must not be broken: (0) a robot may not harm all humanity or, through inaction, allow humanity to come to harm. This would allow a robot to harm individual humans, if so doing prevented an 'existential threat' to all humanity. But how could a robot determine when such a threat exists, and hence killing individual humans to prevent it is permitted?

3.3.3 Fixing Asimov's laws

Other authors have attempted to fix other ambiguities and loopholes in the rules Asimov devised, in order to prevent disastrous scenarios that nonetheless satisfied laws numbered 0-3. For example, Lyuben Dilov [1974] introduced a Fourth Law of Robotics to avoid misunderstandings about what counts as a human and as a robot: (4) a robot must establish its identity as a robot in all cases. This law is sometimes stated as the slightly different: (4´). A robot must know it is a robot. Others [e.g. Harrison, 1989] have also argued for a Fourth Law that requires robots to reproduce, as long as such reproduction does not interfere with laws 1-3.

Asimov's literary exercise was illustrative of a limitation inherent in any rule-based morality: What does the robot do when there are conflicts between the rules? Should rules function as hard restraints? Or can the rules function as guidelines where the system is designed to factor in an array of *prima facie* duties in the actions it considers? Will this open the door to robotic behavior that should be prohibited? Perhaps the biggest challenges confronting designers of rule-based robots or AMAs is how the system will recognize those situations that require application of the rules, and how to ensure that the robot has access to all the information it needs in order to apply rules appropriately. How would a robot programmed with the First Law *know*, for example, that a medic or surgeon welding a knife over a fallen fighter on the battlefield is not about to harm the soldier? The robot would need to understand a great deal about context, exceptions to rules, and human psychology; and its knowledge base would need to be updated regularly.

Roger Clarke [1994] attempted to update and fix Asimov's laws, in what he called "An Extended Set of the Laws of Robotics":

> "The Meta-Law: A robot may not act unless its actions are subject to the Laws of Robotics.
> Law Zero: A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

Law One: A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher-order Law.

Law Two: A robot must obey orders given it by human beings, except where such orders would conflict with a higher-order Law; a robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher-order Law.

Law Three: A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher-order Law; a robot must protect its own existence as long as such protection does not conflict with a higher-order Law.

Law Four: A robot must perform the duties for which it has been programmed, except where that would conflict with a higher-order law.

The Procreation Law: A robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics."

Clarke admits that his revised laws still face serious problems, including the identification of and consultation with stakeholders and how they are affected, as well as issues of quality assurance, liability for harm resulting from either malfunction or proper use, and complaint-handling, dispute-resolution, and enforcement procedures. Our discussion of product liability in section 5 will address many of these concerns.

There are additional problems that occur when moral laws for robots are given in the military context. To begin with, military officers are aware that if codes of conduct or Rules of Engagement are not comprehensive, then proper behavior cannot be assured. One difficulty lies in the fact that as the context gets more complex, it becomes impossible to anticipate all the situations that soldiers will encounter, thus leaving the choice of behavior in many situations up to the best judgment of the soldier. The desirability of placing machines in this situation is a policy decision that is likely to evolve as the technological sophistication of AMAs improves.

Unfortunately, there are yet further problems: most pertinently, even if their glitches could be ironed out, Asimov's laws will remain simply inapplicable to the military context, as it is likely that autonomous military robots will be asked to exercise lethal force upon humans in order to attain mission objectives, thereby violating Asimov's First Law. A further problem, called '*rampancy*', involves the possibility that an autonomous robot could overwrite its own basic programming and substitute its own new goals for the original mission objectives (e.g., the movie *Stealth*). That leads us to a final and apparently conclusive reason why deontological ethics cannot be used for autonomous military robots: it is incompatible with a 'slave morality', as addressed in the following discussion (and further in section 6).

3.3.4 Slavery: A Crucial Problem for Deontology and Robotic ethics

One further problem, specific to robotics, with deontological ethics is the problem of 'slave morality.' Robots in the military would be presumably programmed to follow commands slavishly, and not exhibit anything like true Kantian autonomy. Indeed, the term 'robot' is derived from the Czech word 'robota' that means 'servitude' or 'drudgery' or 'labor' (see 'Appendix A: Definitions'). Such robots could make autonomous choices about the means to carrying out their pre-programmed goals, but not about the goals themselves; they could not choose their own goals for themselves, but they would always be expected to have the goal of obeying orders given by their military commander.

That would collapse (from a deontological perspective) all questions about their ethics into simply questions about the ethics of the military commander, and *mutatis mutandis* for any other use of autonomous robots as slaves. Such an approach would then claim that there is actually no such thing as robot ethics; there are only the ethics of those who command robots. But the concerns with robot ethics crucially concern the consequences of using them—a concern a strict deontological ethics cannot countenance as it insists that one must obey the rules, no matter the consequences. And of course, a key objection (that will affect both deontological and utilitarian ethics) is the plausible skeptical claim that no finite set of rules can ever guarantee ethical behavior in all cases, or at least where the set of possible behaviors is large or practically unlimited. Before addressing that critique, let us examine perhaps the most important objection to deontology (and the resulting alternative approach)—that consequences matter morally, and simply following the rules is morally wrong if it leads to bad outcomes.

3.3.5 Top-Down Approaches: Utilitarian Consequentialism

Utilitarianism represents another attempt to bypass conflicts between rules through an overriding top-down principle that can be applied to all situations. However, with respect to computability, this approach stresses the importance of the outcomes (consequentialism) arising from an action. *Consequentialist* approaches to ethics focus on achieving the best possible outcomes in various situations, and hence typically disdain rigid rules that specify unchanging duties. For example, utilitarianism—the primary consequentialist theory— proposes that an agent should calculate the net consequences arising from the various available courses of action, and then select the action that offers 'the greatest good for the greatest number.' This is a familiar, pragmatic theory in that many policy and business decisions seem to be determined by a weighing of reasons for and against a particular action, and it suggests a simple algorithm for calculating what action one ought to take in a given situation.

However, this approach is not as computationally tractable as it might appear. Practically, there is the calculational objection: it is an impossible demand to calculate the utility of every action; thus, utilitarianism *makes moral evaluation impossible,* as even the short-term consequences of most actions are impossible to accurately forecast, much less the long-term consequences. Problems of how utility might be represented within a computational system, how broadly the consequences of actions should be analyzed, and which agent's welfare should be included in the calculation need to be resolved in order to bring a utilitarian analysis to a successful conclusion. Given limitations of available information, the breadth of variables impinging upon a complex set of interrelated agents, and therefore an inability to accurately predict the consequences of an action, such a calculation poses a tremendous computation load on even the fastest systems. A utilitarian robot may fail to determine which course of action is most acceptable within the time allotted.

But if utility is incalculable, and one's obligation is to maximize utility, much of the theory's value seems to disappear. Worse, there are further objections to utilitarianism: the *absurd implications objection* would, for example, point to some scenario in which a lie is just as moral as truth, if the consequences are the same. Even more fundamental are objections based on (in)justice**.** For example, the *scapegoating* objection would point out that maximizing utility may demand injustice, such as executing an innocent person to prevent a riot that would have resulted in deaths and economic damage. This is to say that utilitarianism, at least in its basic form, cannot readily account for the notion of rights and duties nor moral distinctions between, e.g., killing versus letting die or intended versus merely foreseen deaths (assuming we think such notions and distinctions exist).

Whether deontological or consequentialist/utilitarian, each of the single-principle top-down theories suffers from a version of the frame problem—that is, it requires an impossible computational load due to the requirements for knowledge of the relevant effects of action in the world, the difficulty of estimating the sufficiency of the initial information, and knowledge about the psychology of agents. Nevertheless, humans appear to apply rough and ready top-down evaluations in their selection of courses of action, and so might a robotic system, particularly if the goal is not to create a perfect system but only one that makes better (or just as good) decisions than humans do.

Top-down theories combine strength in defining ethical criteria with a breadth that can be applied to countless challenges. The price of this strength lies in the goals either being defined so vaguely and abstractly that their meaning and their application to specific situations is debatable, or they are defined so rigidly that they fail to produce decisions that are appropriately sensitive to new context.

3.4 Bottom-Up Approaches

The bottom-up approaches to building AMAS are inspired by three sources: (1) the tinkering by

engineers as they optimize system performance, (2) evolution, and (3) learning and development in humans. Bottom-up approaches fall within two broad categories: the assembling of systems with complex faculties out of discrete subsystems, and the emergence of values and patterns of behavior in a more holistic fashion, as in artificial life experiments and connectionist networks.

A variety of discrete subsystems are being developed by computer scientists that could potentially contribute to the building of artificial moral agents. Not all of these subsystems are explicitly designed for moral reasoning. For example, learning algorithms, affective sensors, and social mechanisms might all contribute to the moral acumen of a robot. But computer scientists who wish to build robots with higher-order faculties out of discrete subsystems are confronted with a difficult, and perhaps insurmountable, challenge of assembling components into a functional whole. Whether the aggregation of discrete skill sets will lead to the emergence of higher-order cognitive faculties—including emotional intelligence, moral judgment, and consciousness—can only be known once roboticists go through the exercise of building the systems.

3.4.1 Optimizing Performance

Various trial-and-error techniques are available to engineers for progressively tuning components so that the system approaches or surpasses the performance criteria. Bottom-up approaches to ethics treat normative values as being implicit in the activity of agents rather than explicitly articulated (or even articulatable) in terms of a general theory. Engineers commonly define tasks atheoretically using a performance measure, such as winning chess games, passing the Turing test, walking across a room without stumbling, and so on. Even without a theory of the best way to decompose the task into subtasks, engineers can achieve a high level of performance on many tasks. Sometimes a *post hoc* analysis of the system can produce a theory or specification of how the subtasks yield results. But often the results of such an analysis do not correspond to the kind of decomposition suggested by *a priori* theorizing.

3.4.2 Evolution

Evolution has inspired an array of approaches for developing artificial intelligence from artificial life experiments (Alife) to genetic algorithms and to evolutionary robotics. The theory of evolution has suggested to engineers a model for self-selecting and self-organizing systems that strive toward the optimization of some performance criteria, such as the maximization of profits. The power of evolution is tapped into by selecting those agents, from a collection of similar agents, that are most successful at optimizing a specified fitness (performance) criterion. The selected agents serve as *parents* that are modified and recombined (using a process that is analogous to sexual reproduction) to produce a new generation of agents. This new generation is tested, the best performers selected, and they in turn breed, and so forth. This basic strategy has been successful for producing agents

suited to a wide variety of tasks.

Two ideas contribute to the belief that evolutionary strategies would be helpful for eliciting moral behavior in agents. The first is the contention by game theorists and evolutionary psychologists that moral propensities, such as cooperation and care of the young, may have emerged during the course of evolution are partially encoded in genes and potentially reproducible in simulations of evolution within computer environments. However, as Rodney Brooks has noted, experiments in Alife "have not taken off by themselves in the ways we have come to expect of biological systems" [Brooks, 2002]. The second influencing idea is that optimizing moral performance might be used as the fitness criteria for selecting the best agents. The difficulty with this strategy lies in how the fitness criteria would be represented in a computational system. The slogan 'survival of the most moral' highlights the problem of saying what 'most moral' amounts to in a non-circular (and computationally tractable) fashion.

3.4.3 Learning and Development

Alan Turing was the first to broach the idea that artificial intelligence (AI) should try to mimic child development. In 1950 he wrote: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain" [Turing, 1950].

Jean Piaget, Lawrence Kohlberg, Carol Gilligan and others have proposed developmental theories regarding the way in which children learn about morality [Murray, 2008]. These theories have been adapted into curricula that facilitate the moral development of children. This suggests the possibility that a learning robot might be taken through a similar educational program. However, while machine learning is an important area of research, the algorithms and techniques presently available are not robust enough for such a sophisticated educational project. For the immediate future, machine-learning techniques are likely to be quite rudimentary.

3.4.4 The Value and Limits of Bottom-up Approaches

Individual subsystems can be quite brittle in their performance. However, when integrated successfully, these components can give rise to complex dynamic systems with a range of choices or optional responses to external conditions and pressures. Bottom-up engineering thus holds the promise of a kind of dynamic morality where, as conditions change, the ongoing feedback from different mechanisms facilitates varied responses. It ties into a movement within ethics termed 'particularism', which asserts that no general laws or rules are possible, and each ethical situation is unique.

But the weakness of bottom-up approaches for developing AMAs lies in not knowing which goals to use for evaluating choices and actions as contexts and circumstances change. Bottom-up systems work best when they are directed at achieving one clear goal. When the system has more than one goal, or when the available information is confusing or incomplete, bottom-up engineering is less likely to provide a clear choice or course of action.

3.5 Supra-Rational Faculties

In order to function as moral agents, robots that interact with humans within dynamic multi-agent contexts may require other skill sets in addition to being rational. Which skills the agent will need will vary depending upon the robot's tasks and goals. For example, tasks that socially-viable robots will perform can require emotional intelligence, knowledge about social customs, and the non-verbal cues used to communicate essential information. In addition, the capacity to appreciate and respond to moral challenges may also depend upon the robot having semantic understanding, consciousness, a theory of mind (the ability to deduce the intentions, desires, and goals of others), and perhaps even the capacity to feel pain or empathy. These additional faculties are a tall order for roboticists, although each has already stimulated interesting lines of research that are under way.

Algorithms for reasoning about moral challenges will not lead to appropriate behavioral responses unless the robot has access to the background information describing the situation, the ability to discern which information is essential and which inputs are of ethical concern, and the capacity to recognize inherent and potential conflicts arising from the competing interests of the various agents. In other words, the engineer must determine what the information requirements are for a system making moral decisions. What will the system need to know in order to make an informed decision? What input devices and sensors will it need to get access to this information?

Supra-rational faculties—such as emotions, being embodied in the world, social skills, and consciousness— represent ways that humans get access to essential information that must be factored into ethically-significant choices. Sensory experience and social mechanisms contribute to various refinements of behavior that people expect from each other; they also will be necessary for robots that function to a high degree of competence in social contexts. While robots will not necessarily need to emulate the full array of human faculties, the more sophisticated systems will need mechanisms that provide a similar appreciation of complex social contexts. Furthermore, communicating through facial expressions, gestures, vocal intonation and prosody, and other verbal and non-verbal cues will be helpful in conveying the robot's intentions and facilitating cooperation with humans. This will, in turn, help humans to perceive the robot as being trustworthy.

Given that morally intelligent behavior may require much more than being rational, the challenge of

building AMAs is, from this perspective, a problem of moral psychology, not moral calculation. For designers of morally intelligent systems, the challenge is not how to give them abstract theoretical knowledge but how to develop robots that embody the right tendencies in their reactions to the world and other agents in that world.

3.6 Hybrid Systems

Moral judgment in humans is a hybrid of both (1) bottom-up mechanisms shaped by evolution and learning and (2) top-down criteria for reasoning about ethical challenges. Eventually, we may be able to build morally intelligent robots that maintain the dynamic and flexible morality of bottom-up systems capable of accommodating diverse inputs, while subjecting the evaluation of choices and actions to top-down principles. The prospect of developing virtuous robots offers one venue for considering the integration of top-down, bottom-up, and supra-rational mechanisms.

3.6.1 Virtue Ethics: The Virtuous Robot

There is a foundational critique of all procedural ethics, i.e., any approach that claims morality is determined by following the proper rules. Many contemporary ethicists claim that all procedural ethics fail, because so many theorists have explicitly abandoned the ideas that in ethics: (a) the rules would amount to a decision procedure for determining what the right action would be in any particular case; and (b) the rules would be stated in such terms that any non-virtuous person could understand and apply them correctly.

In robotics, so-called '*friendliness theory*' attempts to deal with this conundrum: rather than using any finite set of top-down rules or laws, intelligent machines should be programmed to be basically altruistic, and then use machine learning in various settings to create a kind of 'best judgment' in how to carry out properly altruistic actions. This approach sidesteps the fundamental calculational and programming problem of how to account for a vast number of unforeseeable eventualities. However, this theory has a problem with military robots: we would not want them to always act altruistically towards some humans. In fact, we would want them to be able to kill the right humans and not the wrong (friendly) ones. Therefore, we need an approach that enables machines to use a basic top-down program plus bottom-up machine learning to be able to function excellently in its military roles and without malfunctioning; it should be fierce towards its enemies, helpful to its allies, and reliable in discerning the difference, including in situations unforeseen by its programmers. What ethical approach can accomplish all this?

Perhaps the most viable hybrid approach that avoids the conflict between duties and consequences, and incorporates both warrior fierceness towards enemies and a gentle kindness towards comrades,

is *virtue ethics*: an approach that sees ethics, not in terms of what rules should be followed, but in terms of what kind of character an agent has—does one have a virtuous character, or is one full of vice? For virtue ethics, morality is not a function of actions but of character. That is, one's actions do not constitute one's morality but rather *reveal* it: ethics is agent-centered, not act-centered. The proper moral question is not "what rule should I follow?", or "what rules apply to this act", but instead "what sort of person will I reveal myself to be (or become) if this is the sort of thing I do?" What would doing this act say about my character?

While virtue theorists differ in their list of virtues, they take their lead from Aristotle in recognizing that the virtues are acquired developmentally through experience and the cultivation of good habits. This emphasis on developing the virtues can be understood as bottom-up learning, while, at least in theory, it is possible to consider the virtues as top-down patterns for evaluating actions programmed into a robot. Because virtue is an 'excellence' and defined in terms of the roles one plays, it is inescapably context-dependent; no single rule or set of rules will be able to dictate what different persons (or robots) in different roles will need to do in different situations. Moral criteria are thereby objective but not categorical rules; instead, they are what Kant called 'hypothetical imperatives' linking good means to good ends [Foot, 1972].

Because the virtues are objectively-beneficial habits for proper functioning, but are role dependent, the objective list of virtues for firemen will be different than the list for auditors or salesmen or soldiers, and so on. Only the most generic virtues—e.g., wisdom, honesty, empathy, justice, etc.—will apply to all social roles. Professional codes are then best understood as list of virtues for a particular social role, rather than a list of rules to follow. Thus, following Solomon [1988], we can say that good eyesight is a virtue in a rifleman; it is a virtue because it helps in achieving the purposes or goals of a rifleman. But while the lack of a properly developed conscience might be thought of as a virtue in a hitman understood narrowly within that particular social role, it is not a virtue in a person simply described as a person, within the all-encompassing set of social roles we call human life. A true virtue is thus an excellence in a role that aids overall human flourishing.

3.6.2 The Top–Down Challenge and the Bottom-Up Approach

The top–down challenge for an engineer designing a robot would be to determine how to represent virtuous patterns and motivations, and how the system would determine which virtue, or which action representing the virtue should be called upon in a particular situation. Given the emotional grounding of virtuous motivations in human beings, a designer of a virtuous AMA might need to decide whether a virtuous machine would also need emotions of its own or some mechanisms functionally similar to human emotions.

The bottom-up approach to implementing virtues in computational systems arises from the

recognition by several theorists [e.g., DeMoss 1998; Churchland, 1995] of the similarity between learning in connectionist networks and Aristotle's discussion, in the *Nicomachean Ethics,* regarding the way in which the virtues are acquired. Connectionist networks provide a bottom-up strategy for building capacities through the recognition of patterns and the building of categories out of complex inputs. Through the gradual accumulation of data, the network develops generalized responses that go beyond the particulars on which it is trained. One difficulty with learned patterns that emerge from connectionist systems is that they are not accompanied by explanation for *why* the action was chosen.

The problems tackled by existing connectionist networks are far from the complex learning tasks associated with moral development. However, the prospect that neural networks might be adapted for some aspects of moral reasoning is an intriguing possibility. Neural networks offer an approach, deserving of attention, for developing robots that embody the right tendencies in their reactions to the world. The bottom-up development of virtuous patterns of behavior might be combined together with a top-down implementation of the virtues as a way of both evaluating the actions and as a vehicle for providing rational explanations of the behavior.

While many technological thresholds must be crossed before the development of a virtuous robot becomes a serious possibility, we believe that this approach to building AMAs should be of particular interest to the military in its long-term planning. A virtuous robot might emulate the kind of character that the armed forces value in their personnel. Furthermore, virtues—deeply rooted in the foundational attitudes and structures of an agent—provide a certain degree of stability, and the prospect that officers can rely upon the performance of the artificial agents they deploy.

3.7 First Conclusions: How Best to Program Ethical Robots

A top-down approach would program rules into the robot and expect the robot to simply obey those rules without change or flexibility. The downside, as we saw with the analogous deontological ethics, is that such rigidity can easily lead to bad consequences when events and situations unforeseen or insufficiently imagined by the programmers occur, causing the robot to perform badly or simply do horrible things, precisely because it is rule-bound.

A bottom-up approach, on the other hand, depends on robust machine learning: like a child, a robot is placed into variegated situations and is expected to learn through trial and error (and feedback) what is and is not appropriate to do. General, universal rules are eschewed. But this too becomes problematic, especially as the robot is introduced to novel situations: it cannot fall back on any rules to guide it beyond the ones it has amassed from its own experience, and if those are insufficient, then it will likely perform poorly as well.

As a result, we defend a hybrid architecture as the preferred model for constructing ethical autonomous robots. Some top-down rules are combined with machine learning to best approximate the ways in which humans actually gain ethical expertise. We humans are hard-wired with various rules through our evolutionary heritage, but these vastly underdetermine actual behavior; hence responsible behavior builds on these underlying (largely unconscious) rules with a healthy dose of indoctrination and peer interaction and all the other types of learning that children do. As a result, a character evolves: a tendency to perform certain roles in the evolving ecology of social life, and either to fail or to perform excellently in those roles.

3.7.1 Further Conclusions on Ethical Theory and Robot: Military Implications

Autonomous robots both on and off the battlefield will need to make choices in the course of fulfilling their missions. Some of those choices will have potentially harmful consequences for humans and other agents worthy of moral consideration. Even though the capacity to make moral judgments can be quite complex, and even though roboticists are far from substantiating the collection of affective and cognitive skills necessary to build AMAs, systems with limited moral decision-making abilities are more desirable than 'ethically blind' systems. The military's comfort with the robots it deploys, and ultimately the comfort of the public, will depend upon a belief that these systems will honor basic human values and norms in their choice of actions. Given the prospect that robotic systems can reduce the loss of personnel during combat, one can presume that the development of autonomous robotic fighting machines will proceed. However, if semi-autonomous and autonomous robotic systems are deployed as lethal weapons, it goes without saying that commanders will need to be confident that the systems will only wield their destructive might on designated targets.

The challenge for the military will reside in preventing the development of lethal robotic systems from outstripping the ability of engineers to assure the safety of these systems. Implementing moral decision-making faculties within robots will proceed slowly. While there are aspects of moral judgment that can be isolated and codified for tightly defined contexts, moral intelligence for autonomous entities is a complex activity dependent on the integration of a broad array of discrete skills. Robots initially will be built to perform specified tasks. However, as computer scientists learn to build more sophisticated systems that can analyze and accommodate the moral challenges posed by new contexts, autonomous robots can and will be deployed for a broad array of military applications. So for the foreseeable future and as a more reasonable goal, it seems best to attempt to program a virtuous *partial* character into a robot and ensure it *only enters situations in which its character can function appropriately*.

Theorists continue to debate whether strong artificial intelligence is possible [e.g., Searle, 1980;

Russell and Norvig, 2003]. However, even AI systems with more limited intelligence will require some degree of moral sensitivity in the choices and actions they take: "If there are limitations in the extent to which scientists can implement moral decision making capabilities in AI, it is incumbent to recognize those limitations, so that military planners do not rely inappropriately on artificial decision makers" [Wallach et al., 2008].

For military robots, that virtuous character will likely involve ensuring that the LOW and ROE are programmed in (which may differ from mission to mission) and steadfastly obeyed, as a proxy for a full-fledged morality. Such an approach has several advantages. First, any problems from moral particularism or other problems with general ethical principles (including misguided moral relativism) are skirted. Second, the relationship of morality to legality—a minefield for ethics—is likewise largely avoided; the LOW and ROE make clear what actions are legal and illegal for robots; and for military situations, that can serve as a reasonable approximation to the moral-immoral distinction. So the background of ethics for autonomous robots in the military can, at least for now, become part of our discussion about the programmability of the LOW and ROE. What this means for how we should program ethical robots, and the implications of this approach for the Laws of War and the ethics of risk, is examined in the next section of this report.

4. The Laws of War and Rules of Engagement

For any of the ethical frameworks we have identified in the previous section—deontological ethics, consequentialism/utilitarianism, virtue ethics—the current state of robotics programming (the AI or control software in robots) is not yet robust enough to fully accommodate them. Nevertheless, understanding those ethical theories now is essential for illuminating a thoughtfully-planned path with respect to developing ethical behavior in robots. In the meantime, for military robots, a reasonable proxy for any such ethical theory seems to be found in the Laws of War (LOW) and Rules of Engagement (ROE)—an alternative programming approach with several advantages, as explained at the end of the last section; but, as we explain here, it also has its shortcomings. In this section, we turn to the LOW and ROE, including their relation to just-war theory, and their suitability as an interim programming solution.

4.1 Coercion and the LOW

To understand the LOW and ROE, and to evaluate their viability as a programming approach, we must first understand their origins in just-war theory and, even more basic, the nature of warfare, i.e., what do we need LOW and ROE in first place, and why can't we say 'anything goes' in war?

War, however regrettable, has been an inescapable aspect of human life to date. Autonomous robots have the capacity to radically change the nature of war, and perhaps even eventually lead to its cessation. But during the 'growing pains' of robotics development, our autonomous machines and systems could make the horrors of war either much better or much worse; hence, the ethics of robotic war will be one of the most important subjects of the next decades of the future. To understand the nature of war, we can see it as a type of forcible coercion that nations engage in as a means of attaining their political goals. Let us define the term as follows:

Coercion. *The use of force and or violence, or the threat thereof (i.e., intimidation), in order to persuade*. Coercion is a sad reality both within societies and in international affairs. Within recognized societies, legitimate coercion is exercised by the state, through its police and judiciary, to help restrain and deter illegitimate coercion by individuals. But in the international arena, no supra-national institution has clear and effective coercive power over nation-states who perform illegitimately coercive acts. Hence, states must resolve issues of illegitimate coercion amongst themselves, often through coercion of their own. Hence, we have the phenomenon of war: armed

conflict between states, which attempts to coerce some desired outcome in lieu of other means (e.g., negotiations) of attaining international agreement.

Naturally, if states are engaged in legitimate forcible coercion in order to deter or punish illegitimate coercion, then we must have some agreed upon means of distinguishing legitimate from illegitimate coercion amongst states in war. This is called 'just-war theory', and it attempts to spell out when beginning the coercion of war is morally legitimate and when it is not (termed *jus ad bellum*); and further, what means of wartime coercion are morally permitted (*jus in bello*) and what one should do in the aftermath of officially ending such coercion (*jus post bellum*).

4.2 Just-War Theory and the LOW

So, the Laws of War, also known as the Laws of Armed Conflict (LOAC), concern the legal and moral legitimacy of practices that nations engage in during the interstate forcible coercion that we call 'war.' As mentioned, the Laws of War are divided into three basic categories, with the first two being of general and long-standing acceptance, but the third forming a relatively new emphasis, albeit of increasing import in contemporary asymmetric and non-state warfare:

1. *Jus ad bellum*: Law concerning acceptable justifications to use armed force and declare war.
2. *Jus in bello*: Law concerning acceptable conduct in war, once it has begun.
3. *Jus post bellum*: Law concerning acceptable conduct following the official or declared end of a war (including occupations and indefinite ceasefires, the acceptance of surrender, and the treatment and release of prisoners of war (POWs) and enemy (non-)combatants after conflict has officially ceased).

These three categories are typically (but not always) asserted to be independent; so, the morality and legality of a state deciding to go to war (*jus ad bellum*)—typically a political decision made by a state's political leadership—have long been considered independent of the morality and legality of one's actions in waging war (*jus in bello*); the latter is typically the province of a state's professional military, not its political leadership. For instance, we might hold that an American soldier who participated in the My Lai massacre was guilty of war crimes (a violation of *jus in bello*), but not because the Vietnam War was itself unjust (even assuming that armed conflict was a violation of *jus ad bellum*). Likewise, one might have fought a just war and done so in a just fashion, but it may still impose unjust conditions on the vanquished, thus violating *jus post bellum*.

4.3 Just-War Theory: *Jus ad Bellum*

A traditional emphasis of just-war theory concerns when it is morally acceptable for a state to begin or participate in the extreme forcible coercion that is war, that is, *jus ad bellum.* The ancient Greeks (Aristotle) and Romans (such as Augustine) considered these issues, but the natural law tradition associated with Aquinas began the systematic consideration of the issue. Natural law and social contract theorists have continued it in the work of such luminaries as Vitoria, Grotius, Locke, and Kant; the 20th century adapted this just-war tradition in gradually creating the internationally accepted LOW. Hence, a brief explanation of just-war theory is needed; current influential theorists include Walzer, Orend, and O'Brien, with a rough consensus on the following as necessary conditions for moral *jus ad bellum*:

a. *Proper authority*. War must be waged by a competent authority (normally an internationally recognized state) for a publicly stated purpose, i.e., 'secret wars' are immoral. But this poses a possible dilemma: Would then all revolutions be immoral? This requirement of *jus ad bellum* has considerable difficulty in defining any non-state rebellions (including popular revolutionary movements) as moral. Compare this to the problem of distinguishing between 'freedom fighters', terrorists, and mere criminal behavior.

b. *Just cause*. There must be sufficient and acceptable justifications for entering war. Specifically, the war must be in self-defense, against unjust attacks on innocent citizens or states, to restore rights wrongfully denied and to re-establish a just order (against unjust rebellion). When a state has forfeited its moral right to govern its people—so it is no longer a 'minimally-just state'—other nations may invade it in order to carry out humanitarian interventions in the self-defense of its people, e.g., in Kosovo or Darfur. The state, no longer being minimally just, has forfeited its own right to self-defense. Problem: In addition to the obvious challenge of determining when a state is no longer 'minimally just', developments in non-state warfare, especially terrorism, complicate this requirement.

   However, offensive wars may be justified if to enforce justice for oneself. Problem: The so-called 'Bush Doctrine' and other policies in modern warfare that justify a preemptive war against looming threats would fail the usual interpretation of having a just cause for war. Current scholarship thus focuses on the proper interpretation of self-defense against a merely potential, but not actual, threat; common criteria include the imminence and seriousness of the threat. (See section 6 on risk assessment for more.)

3. *Proportionality*. The good achieved by war must be proportional to the evil of waging it. Therefore, it is immoral to wage a massive war to remedy a small wrong (e.g., the 'Soccer War' of 1969 between Honduras and El Salvador).

4. *Last resort*. Peaceful means of avoiding war have been exhausted, e.g., negotiations must have been tried and have failed; thus, war is acknowledged as a last resort. Problem: This again makes any so-called pre-emptive war problematic—after all, how can one side be sure that negotiations have completely failed, until the other side actually attacks?

5. *Reasonable success.* This requirement asserts that there is no point fighting a war one cannot possibly win. Because the cost of war is so terrible, it is immoral to fight by futile coercion with no possibility of attaining one's goals, since that would lead to unnecessary, useless casualties; so one must resist in some other way.

6. *Right intention*. Finally, there must be a subjective as well as objective moral rightness in entering a war. One must have the morally-correct motivation and mindset in engaging in war, rather than illegitimate goals (e.g., revenge or economic gain) in mind. For example, to attack the enemy in self-defense, with the intent to merely gain victory and (a lasting?) peace, would fit the requirement of right intention; to perform exactly the same actions, but with the mindset of merely satisfying a violent bloodlust, or gaining control of valuable properties such as oil, would fail this requirement.

4.3.1 Robots and Jus ad Bellum

Peter Asaro [2007] raises an objection to the use of robots in war, that the development of military robots seems to fail a *jus ad bellum* test, because they would embolden political leaders to wage war; robotic soldiers would lower barriers to entering a war, since they would reduce casualties among human soldiers and therefore also a significant political cost. Relatedly, Sparrow [2007] and Sharkey [2007a] object to the wartime use of robots, because they would make war (more) risk-free, at least on the deploying side, but war morally requires there be a terrible cost so that political leaders do not choose it so casually.

Note that this argument is indirect: no one seriously contends the robots themselves, particularly if programmed with a suitable 'slave morality', will themselves be directly effecting *jus ad bellum* violations. Rather, autonomous robots, with their promise of fewer human casualties, will make war less terrible and therefore more tempting, plausibly enticing political leaders to wage war more readily. But such an argument has multiple flaws. First, to claim that robots have bad consequences for declaring war is a consideration that would be handled by the non-consequentialist requirements for declaring a just war: using robots or not makes no difference as to whether the war is (a) in self-defense, (b) proportionally achieving a good greater than the evil of war, (c) a last resort, and so forth.

Second, if any technology (from better armor to longer-range missiles) makes it easier to enter a war to the extent that it reduces risks on our side, these objections seem to imply that we should not make any improvements in the way we prosecute a war and, indeed, should return to more brutal methods (e.g., bayonets). But surely this is ridiculous or, at the least, counterintuitive. Indeed, the increasing horrors of war have reinforced the need for *jus ad bellum* and *jus in bello restrictions*, not undermined them. It is likely the advent of military robots will cause further sophistication in such just-war considerations and make war ever more ethically waged, as is indeed a goal of this report.

Further, we can acknowledge that war is terrible and ought to be avoided whenever morally possible, but at the same time we can adopt a 'deterrence' strategy to avoid war: to create such an overwhelmingly-powerful military force that no one would want to risk a war with us. Granted, this may be an unrealistic goal and may merely spark an arms race; but (so far) this approach seems to be working reasonably well with nuclear weapons—which suggests that the dream of a 'risk-free war' is unrealistic as well, and any lowering of barriers to war may be temporary at best, if even significant. Therefore, the dream of incurring no friendly casualties in war still remains ever elusive, even if robots are deployed on the frontlines first. (We will discuss this and other objections further in section 7.)

4.4 Just -War Theory: *Jus in Bello*

There are serious issues with traditional *jus ad bellum*, and the doctrine will continue to evolve as the technology and asymmetric nature of contemporary warfare change. But because this report concentrates on robotic ethics, and especially the ethics of deploying autonomous robots by the military, *jus ad bellum* issues will herein be dealt with only insofar as they affect the *jus in bello* use of robots. It is exceedingly unlikely in the near- or even medium-term that robots will be in any way responsible for declaring war or even inadvertently starting a war, and the moral use of robots in *jus post bellum* situations will largely flow from the morality of using them *in bello*. Hence, we focus now on the LOW and ROE for *jus in bello*, especially with respect to the use of autonomous robots.

4.4.1 Total War Doctrine: Is There Really a Jus in Bello*?*

"War is hell", reportedly said US by General Sherman—and he destroyed infrastructure and burned to the ground the cities and farms of civilians in Georgia on his march to the sea [Davis, 1980]. Sherman believed that, given the just cause he had in waging war, he was permitted nearly any means to victory, including intentionally harming civilians. By World War 2, this view became known as 'total war' doctrine, espoused by those who saw nothing wrong with launching V-2 rockets on the citizens of London, or firebombing the citizens of Dresden or Tokyo, or dropping nuclear weapons on Hiroshima and Nagasaki. This view asserts that, assuming *jus ad bellum* is satisfied, there are no *jus*

in bello restrictions. That is, one may do whatever is needed to win the victory in a just war, in whatever way one sees fit; our enemies have forfeited their right to any consideration by unjustly beginning their forcible coercion, and deserve whatever they get.

The defenders of total war doctrine, as well as certain 'realist' interpretations of state sovereignty and action, sometimes defend their view by taking the actual state of war to be the absence of any moral or legal structure or standing. Accordingly, they regard the LOW as an elaborate public relations fantasy that nations sometimes use when it suits their *Realpolitik*, or (less cynically) as simply a misconceived enterprise without any actual theoretical or practical grounding. War, these realists claim, is an inherently amoral enterprise, and the laws of the state no longer apply to those waging war against the state, as they have rejected any social contract to abide by civil behavior; hence, there is no basis for any moral or legal code concerning warfare. Laws and morality, it is claimed, are only possible with a settled nation-state, in the absence of war and with the expectation of the rule of law; an attempt to understand the morality or legality of war is then to attempt the oxymoronic.

On this view, war (unlike usual criminal activity) occurs against the background of a complete absence of normal law and order; hence, it becomes absurd to define 'war crimes' as if they constitute a violation of the proper conduct of war operations, in analogy with how normal 'crime' violates the proper functioning of a peaceful civilian society. War is not merely a legally-defined 'business by other means' but instead is a last resort of a sovereign state whose peaceful political life (which is the background for 'normal' crime) is at risk. Total war thus eliminates the usual *jus in bello* distinction between combatants and noncombatants, seeing all those who help the opposing state (or merely reside within it) as legitimate targets in a nation's existential struggle. Total war thus undercuts the applicability of just-war theory to the actual conduct of war: it denies the possibility of any obligatory *jus in bello* restrictions.

But this view appears both unrealistic and morally indefensible. While it is true that the international arena is not yet sufficiently similar to a well-governed state such that wars are simply considered to be international crimes and soldiering is merely international police work, it is also true that international relations are hardly a Hobbesian 'war of all against all' [Hobbes, 1651]. As already alluded to, a rich history of customary international law has been gradually built up and accepted by warring parties through the ages, and international institutions have gradually come to exist which can enforce them. Throughout history, as a matter of honor, prudence, strategic foresight, or even mercy, there have been *jus in bello* restrictions that acquired both moral and legal weight.

This trend toward seeing war as an activity with rules or virtues that sanction proper and improper behavior has only gained strength as states have acquired an institutional professional military,

especially one independent of those making *jus ad bellum* decisions. This is the case even when such professional militaries are voluntarily joined, for most if not all of the individual soldiers no longer have a meaningful right to refuse to fight wars that they find morally objectionable, nor do they have the moral right to fight wars in any way they see fit. Instead, professional soldiers have a code of conduct that details their proper and improper functioning in their various roles, just as other professions do. They cannot be meaningfully held responsible for decisions by politicians over which they have no control; but they can be held responsible for performing their roles in war in a way the international community recognizes as legitimate and avoiding illegitimate means of performing those roles.

As a result, and despite (or because of) the movement toward total war in World War 2 through indiscriminate weapons of mass destruction, arguments for total war doctrine now have a sense of the ridiculous. The Geneva and Hague Conventions have delineated various *jus ad bellum* restrictions on war ever since 1864, with major protocols added in 1977 and amendments continuing to be debated and accepted up to the present. The international community and international law have thus come to a widespread consensus that total war is immoral and cannot possibly be justified. While morally waging war does legitimate the killing of those who are waging war for the enemy, it does not legitimate mass murder (unjustified killing) of non-combatants, or worse; and there are indeed worse things than death. War is now both too dangerous and too professionalized to be fought so cavalierly, without rules or restrictions.

4.4.2 Traditional Jus in Bello

Total war is thus morally unacceptable; there must be *jus in bello* restrictions for a war to be morally fought, which reflect the virtues of a morally-just warrior. Such a 'warrior ethos' [Oh, 2008] is widely accepted among the professional military. Just-war theory thus demands a "*fundamental moral consistency between means and ends* with regard to wartime behavior" [Orend, 2006, p. 105]. As just wars are limited wars, not total wars, there will be restraints on the means of permissible wartime coercion. And as robots themselves will not be declaring war for the foreseeable future, the direct relevance of just-war theory for autonomous military robots will deal with how they would conduct themselves in prosecuting military activities—and so the relevant issue is *jus in bello,* divided into external rules (how a state's military treats its enemies) and internal rules (how a state's military treats its own people).

Much of the just-war tradition [e.g., O'Brien, 1981] asserts only two basic necessary conditions for the external rules of *jus in bello*:

1. *Proportionality.* Again, the military ends must be proportionate to the means: no unnecessary violence is to be used in order to attain one's military goal, but only a level of

force proportionate to attaining one's goal. To drive the enemy from a hillside, artillery shells may be used; a nuclear weapon that obliterates the hillside and all other life within 100 square kilometers must not be used, as it would be wildly disproportionate. Robots would need to learn how to apply force proportionate to their goal, using some operational program that involved properly computing the minimal force necessary for military success, i.e., using the accepted military criteria of 'military necessity.' After testing, it is easy to imagine that robots could perform at least as well as humans in deploying no greater violent force than needed, and thereby passing the 'military Turing test' for moral deployment.

Under proportionality, Walzer and others also include other aspects of traditional *jus in bello* that reject any means '*mala in se*'—that is, evil in themselves—because they violate human rights whenever used, such as rape [Orend, 2001, p. 124]. Robots presumably can be easily programmed to avoid such means. Proportionality also informs the moral treatment of POWs, such as 'benevolent quarantine': POWs may be stripped of weapons, isolated from fighting, and questioned; but there remains the moral requirement not to torture, beat, starve, or medically experiment upon POWs, as agreed upon in the Hague and Geneva Conventions. Whether or not all such protections apply to 'enemy combatants' in the so-called 'War on Terror' is a matter of political discussion; in any case, even the current US administration does not suggest that there are no restrictions on the treatment of 'enemy combatant' prisoners. Whatever the Laws of War amount to in these cases, programming robots to obey them poses no special problems over and above the basic problem of robot discrimination and classification of humans into their proper *jus in bello* categories, and then meting out the appropriate treatment. Thus, we see the next requirement may be trickier for robots.

2. *Discrimination and non-combatant immunity.* One must attempt to discriminate between combatants and noncombatants (civilians), and noncombatants must not be intentionally killed. By engaging in warfare, enemy soldiers become legitimate targets of lethal force in order to coerce their surrender and thus end their resistance to your victory; but those who are not combatants do not forcibly oppose one's goals in war and do not impede victory directly. Hence, as they need not be forcibly coerced in order to attain victory, it is immoral to do so. In short, if someone is not directly engaged in intentionally harming you, it is morally impermissible to intentionally harm them, sometimes termed the 'principle of self-defense.' Hence, we can see that *jus in bello* prohibits weapons that are intrinsically disproportionate, such as thermonuclear weapons in conventional wars, or those that fail to discriminate between combatants and civilians, such as most biological or chemical weapons—and perhaps even many modes of 'cyberattacks' on computer networks [Rowe, 2008].

4.4.3 The Doctrine of the Double Effect (DDE)

We should note well that the requirement of civilian immunity is *merely* that noncombatants must not be intentionally killed or harmed, not that they must not be harmed at all. The latter requirement in practice would lead directly to pacifism, as no war yet fought or practically imagined could guarantee a complete absence of civilian casualties. But if noncombatants can never be legitimate targets, how can it be morally legitimate to harm and even kill them? The usual way out of this problem of 'collateral damage'—that in practice, all those who wage war foresee that some noncombatants will inevitably be harmed—is to use a time-honored ethical principle (originating from natural law ethics) called the Doctrine of the Double Effect (DDE) that is a central principle in both the LOW and the specific ROE that the military specifies for each mission, as defined:

Doctrine of the Double Effect. In the DDE, an action may be morally permissible, even if it is foreseen that it will cause a bad effect, if certain conditions are met [McIntyre, 2004]:

a. The act itself is not morally wrong (e.g., killing combatants in wartime);
b. The good effect is produced directly by the action, and not by the bad effect (e.g., winning is produced by killing of the enemy combatants, not by terrorizing or murdering civilians; the use of nuclear weapons or widespread chemical/ biological dispersal (as in terrorism) fail this criterion);
c. The good effect is sufficiently desirable to compensate for allowing the bad effect (winning is worth killing civilians); and,
d. The bad effect must not be intended, but merely foreseen and permitted (e.g., we would be happy if all Iraqi civilians escape, but alas, one foresees they all will not, and our weapons never intentionally target them).

According to the DDE**,** one can kill noncombatants only if the intention of the actor is good, that is, his or her aim is narrowly at the intended effect; the 'evil' effect is not the goal, nor a means to the goal; and the warrior seeks to minimize evil involved, making any evil unintentional. That is, one can engage in military actions that one foresees will result in an evil consequence (such as harming noncombatants) as long as that harm was not intended and one attempts, as best as one can, to minimize the unintended harmful consequences. 'Military necessity' thus permits collateral damage, as long as it was either unforeseen, or foreseen but unintended and necessary to the attainment of the military goal or objective. As a more difficult application of the DDE, consider the following: May we morally target the opposition's military bases? On one hand, it seems not, since noncombatants work there (e.g., doctors, cooks, janitors); but by the DDE, it may be permissible, as long as the noncombatants are not targeted. Therefore, we should not attack non-combatant sleeping quarters and perhaps time our attack during a period in which a minimum number of non-combatants occupy

the site (e.g., late at night), making any resulting non-combatant deaths the accidental casualties of taking out one's intended military (combatant) target.

4.4.4 The Principle of the Double Intention (PDI)

Walzer and many others in the just-war tradition are also focused on clarifying the DDE so to make clear that it is illegitimate in a just war to intend harm to noncombatants. Arkin [2007] thus appropriates an aspect of Walzer's work in his work on devising methods of programming ethical autonomous robots, and in particular endorses Walzer's version of the DDE: the Principle of the Double Intention (PDI) which is essentially the DDE plus a further ('double') intention that combatants are not only to refrain from intending harm to civilians, but they are also to take precautions to reduce risk to civilians, even at the expense of increasing risk to themselves.

Immediate questions for ethics are raised by the PDI: For example, what does it mean to intend to reduce civilian risk, and how much should civilian risk be reduced [Lee, 2004]? For instance, in the technologically-asymmetric warfare typical of America's military actions in Iraq and Afghanistan, are long-range precision-guided munitions, which allow accurate targeting to within a few feet, morally allowed by the principle of discrimination—or are soldiers morally obliged to engage in close quarters combat (at far greater personal risk) in order to further minimize the possibility of civilian harm? Walzer's PDI gives no clear answer, unfortunately. The principle Walzer appeals to is one from liability law—the 'principle of due care'—that is, that one exercised due care (including potentially creating some risk to oneself) before targeting the enemy, and hence did not heedlessly attack civilians. An example would be requiring soldiers to move in closer to a target to ensure they hit it and not nearby civilians, even at some risk to themselves. But how close is morally mandated? And what of bombers flying sorties with 'smart' weapons, who can fly higher and farther away (and hence more safely) and still hit their targets reliably—but how reliably? What level of reliability and accuracy constitutes 'due care'? (We will discuss liability law further in the next section of this report.)

What is clear is that Walzer argues that even in war, moral agents must minimize the foreseen harm (to the undeserving), even if this will involve accepting additional risk or foregoing some benefit. The potential breakthrough that robots present here is trumpeted by Arkin, who believes it is possible to create robots that will do better than human soldiers at satisfying Walzer's additional condition in the PDI—which is especially difficult for humans who understandably are reluctant to minimize foreseen harm to others at the cost of a greater risk to their own life, but this should be easier for literally selfless robots who do not prioritize their own continued existence over obeying their ethical programming.

For current tele-operated military robots, such as the Predator UAV, the current understanding of the requirement of discrimination involves the need for 'eyes on target': the weapon cannot fire until and unless the human tele-operator has the target firmly acquired in its sights, and no civilians are in the bullseye. But the time lag between remotely pulling the trigger and the weapon actually firing, along with all the vulnerabilities in the electromechanical connections in between, mean that eventually a robot with real-time decision-making capability—a sufficiently autonomous robot—should be able to do as well or better than a human operator in such discrimination. Closer to the target, the robot likely would be more effective in preventing unintended deaths. At that point, it seems *jus in bello* would permit or even demand that such autonomous robots be used, and the requirement of *human* eyes on target—i.e., that robots be tele-operated—would be morally scrapped, as the best means of employing the principle of discriminating between combatant and non-combatant targets can then be done by a machine. We already accept that, due to gravitational forces, computers can fly in situations that humans cannot; it is plausible they will soon make better and more moral targeting decisions as well.

4.5 Rules of Engagement and the Laws of War

The Rules of Engagement comprise directives issued by competent military authorities that delineate both the circumstances and the restraints under which combat with opposing forces is joined. For robots, the specific ROE for each mission will have to be programmed in (which may raise technical issues), but there are no special ethical concerns with the ROE, as long as they do not violate the already extant *jus in bello* restrictions of the LOW. Hence, the ROE would constitute an additional ethical issue for the morality of deploying military robots *only if* the competent military authorities were to program in a ROE that violated the underlying LOW. While certainly possible, this raises no culpability issues that do not already exist with human soldiers. For instance, a 'loosened' ROE that permits cross-border attacks into sovereign nations with which we are *not* formally at war, in cases where our troops witness attacks originating from the region and even if those attacks are not directed at us, arguably violates the LOW, specifically the self-defense requirement.

If this or any other ROE does violate the LOW, the ethical result of using robots may be a moral improvement, since robots properly programmed to never violate the LOW would refuse to follow immoral orders, unlike human soldiers who are trained to unfailingly follow all orders. With robots, we may be better positioned to ensure compliance with the *jus in bello* aspects of the LOW, which is a substantial argument in favor of deploying such robots. (We return to the issue of a robot disobeying an order in section 7.)

4.6 Just-War Theory: *Jus post Bellum*

President George W. Bush declared the end of major combat operations for coalition forces in Iraq only a couple of months after hostilities began; yet the insurgency ever since has caused far more casualties than the actual war against the Iraqi government ever did. It thus seems that robots, with suitable *jus post bellum* programming, could also serve as peacekeepers without either the casualties or tendency to target civilians that are among the problems of using human troops in peacekeeping roles. But one objection to robot peacekeepers is that having machines occupy some city or patrol the streets won't help win the hearts and minds of the occupied or vanquished. Could robots be so off-putting, so overwhelming or offensive, that they make lasting peace more difficult to achieve?

This will again be a concern that we return to in section 7 of this report, but one may initially and reasonably expect that, as local populations gain experience with robot peacekeeping that routinely performs in a morally equivalent or superior way to human peacekeeping, their worries will soon ease. After all, robots presumably will not be raping, pillaging, degrading, taunting, or stealing food from the local population, as might occur with human peacekeepers fueled by adrenaline, emotions, and perhaps hatred. And improvements in the appearance of robot peacekeepers may also do much to assuage this worry; just as robotic lethal weaponry is often made to look fearsome in order to strike terror into enemy forces, robotic peacekeepers could be designed to appear friendly and non-threatening.

4.7 First Conclusions: Relevance to Robots

In the not-too-distant future, relatively autonomous robots may be capable of conducting warfare in a way that matches or exceeds the traditional *jus in bello* morality of a human soldier. With a properly programmed slave morality, a robot can ensure it will not violate the LOW or ROE, and it can even become a superior peacekeeper after official hostilities have ceased. And of course, having robots fight for us promises to dramatically reduce casualties on our side and may become a fearsome enough weapon that eventually war will cease to be a desirable option by nation-states as a means of resolving their differences. Once such robots exist and have been properly trained through simulations, there will be little moral justification to keep them sidelined: if war is to be fought, we will have good moral reason to have the robots do the fighting for us.

In the meantime, there are still a number of concerns to address related to risk, ethics, and technical challenges. In the next sections, we will continue a discussion on legal liability and responsibility, as well as a broader discussion about technology risk assessment in military robotics.

5. Law and Responsibility

The use of robots, particular military robots with the capacity to deliberately do harm and which have increasing degrees of autonomy, naturally raises issues with respect to established law and liability. Assuming we can program morality into robots in the first place, using military law—i.e., Rules of Engagement—as a behavioral framework seems to be reasonable or at least more manageable than attempting to program in the much larger set of society's civil and criminal laws. But what would happen if a robot commits some act outside the bounds of its programming—who then becomes responsible for that action?

The answer perhaps depend on the cause, whether the act results from a programming error or malfunction or accident or intentional misuse. But in any case, we would be hard-pressed to assign blame *today* to our machines; yet as robots become more autonomous, a case could be made to treat robots as culpable legal agents. This section investigates several issues related to legal responsibility and robots, both current and future.[5]

5.1 Robots as Legal Quasi-Agents

How might the law treat robots as potential legal agents? There are several relevant aspects of the law that might bear upon robots, and we will consider each in turn, after a brief overview. In the most straightforward sense, the law has a highly developed set of cases and principles that apply to *product liability*, and we can apply these to the treatment of robots as commercial products. As robots begin to approach more sophisticated human-like performances, it seems likely that they might be treated as *quasi-agents* or quasi-persons by the law, enjoying only partial rights and duties.

A closely related concept is that of *diminished responsibility*, in which agents are considered as being not fully responsible for their own actions. This will bring us to the more abstract concept of *agency* itself in the law, and how responsibility is transferred to different agents. Finally, we will consider *corporate punishment*, which is relevant both as it applies to cases of wrongdoing in product liability, but also because it addresses the problem of legal punishments aimed at non-human agents, namely corporations.

[5] We thank and credit Peter Asaro for his contribution to the discussion here.

5.1.1 Responsibility and Liability: Robots as Products

In the system of Anglo-American law, a distinction is drawn between criminal and civil law. Civil law is traditionally called tort law and deals primarily with property rights and infringements, such as damage to property or other harms, and seeks justice by compelling wrongdoers to compensate those who were harmed for their loss. Criminal law deals with what we often think of as moral wrongdoing, stealing, murder, etc., and seeks justice by punishing the wrongdoer. The difference is that between someone building a toy robot which shoots little plastic missiles that causes several small children to choke to death, and someone who builds a robot with a built-in bomb that kills a number of people on a public street. In each case there is a robot causing death, but in the first case the parents of the children would file a lawsuit against the manufacturer seeking monetary compensation, and in the second case the government would find, arrest, prosecute and punish the individuals responsible. Let us set criminal law aside for the moment, however, as civil law appears more relevant to robots as they now exist, insofar as they might be capable of material wrongdoing.

Even if we make no assumptions about the intentions, consciousness, or moral agency of robots, we can still apply the basics of civil law to robots as they now exist. That is, we can assume that robots are completely unremarkable technological artifacts, no different than toasters or cars, and there are still legal and moral issues connected with their production and use. In fact, the companies that currently manufacture robots, such as the Furby™ and AIBO™, can be held accountable under these laws, and therefore almost certainly employ and retain lawyers who are paid to advise them on their legal responsibilities in producing, advertising, and selling these robots to the general public. Furthermore, it seems that many of the concerns about the possible harms that robots might cause would ultimately fall under this mundane interpretation. While these may not be the most philosophically-challenging issues regarding robot ethics, they seem likely to be the most common.

5.1.2 Standards of Liability

The relevant legal concept in cases like our toy robot that chokes small children is *negligence*. Negligence implies that the manufacturer failed to do something that was morally or legally required, and thus they can be held responsible for certain harms produced by their product—in legal terminology this is called *reasonable care*. Legally culpable forms of negligence in product liability cases depend upon either *failures to warn*, or *failures to take proper care*. A *failure to warn* occurs when the manufacturer was knowingly aware of a risk or danger but failed to notify consumers of this risk. This is the reason why there are now so many warning labels on various products, and in the example above the manufacturer might avoid liability by putting a label on the package stating that the robot contains parts that are a choking hazard to young children. A *failure to take proper care or avoid foreseeable risks* is more difficult to prove in court because it is more

abstract, and involves cases where the manufacturer cannot be shown to have known about a risk or danger from the product. In these cases, it is argued that the given danger or risk was in some sense obvious or easily foreseeable, even if the manufacturer failed to recognize it. In order to prove this, the plaintiff's lawyers often bring in experts to testify that those risks were obvious, and so forth.

Another interesting aspect of liability is that it can be differentially apportioned. That is to say, for example, one party might be 10% responsible, while another is 90% responsible for some harmful event. This kind of analysis of the causal chains resulting in harms is not uncommon, especially in traffic accidents and product liability cases. In many jurisdictions there are laws imposing joint and several liability, which holds all parties equally responsible for compensation, even if they are not equally responsible for the harm. Nonetheless, these cases still recognize that various factors and parties contribute differentially to some event.

Differential apportionment could be a useful tool when considering issues in robot ethics. For instance, a badly-designed object recognition algorithm might be responsible for some damage caused by a robot, but a bad camera could also contribute, as could a weak battery, or a malfunctioning actuator, and so on. This implies that engineers need to think carefully about how the subsystem they are working on could interact with other subsystems—whether as designed or in unintentional partial breakdown situations—in potentially harmful ways.

Further, the context in which the robot has been placed, such as the instructions given by its owners, may also be the principle, or contributing, cause of some harm in which a robot is the proximate cause. In short, there is a limit to what robot engineers and designers can do to limit the potential uses and harms caused by their products, because other parties (i.e., the consumers and users of robots) will choose to do all sorts of things with such products and will have to assume the responsibility for those choices. Similarly, there will always be risks inherent in the use of robots, and at some point the users may be judged by a court to have knowingly assumed these risks in the very act of choosing to use a robot.

The potential *failure to take proper care*, and the reciprocal responsibility to take proper care, is perhaps the central issue in practical robot ethics. What constitutes proper care, and what risks might be foreseeable, or in principle unforeseeable, is a deep and vexing problem. This is due to the inherent complexity of potential future interactions and the relative autonomy of the product once it is produced. Sophisticated robots that will be capable of interacting with people and the world in highly complex ways, and that may develop and learn new ways of acting which extend beyond their intended design, present a difficult future in which to foresee risks. Robot ethics shares this double-edged problem with the bio-engineering ethics—both the difficulty in predicting the future interactions of a product when the full scope of possible interactions can at best only be estimated,

and in producing a product that is an intrinsically dynamic and evolving system whose behavior may not be easily guided after it has been produced.

The classic defense against charges of *failures to warn* and *failures to take proper care* is the *industry standard defense*. The basic argument of the *industry standard defense* is that the manufacturer acted in accordance with the stated or unstated standards of their industry. Thus, they were merely doing what other similar manufacturers were doing and, in doing so, taking proper care as measured against their peers. This need for a relative measure again points to the vagueness of the concept of proper care, and the inherent difficulty of determining what specific and practical legal and moral duties follow from the obligation to take proper care. This kind of defense also fails to tell us what sorts of practices *should* be followed in the design of robots. That is, robot ethics should be concerned with the establishment of standards for the robot industry which will ensure that the relevant forms of proper care are taken. It seems that this ought to be one of the industry's top objectives for future research, and there is quite a bit more to be said about this issue; but for now we will stay with the law in our discussion.

5.2 Agents, Quasi-Agents, and Diminished Responsibility

The law offers several ways of thinking about the distribution of responsibility in complex cases. As we saw in the previous section, responsibility for a single event can be divided amongst several parties, and each party can even be given a quantitative share of the total responsibility. We will now consider how even a single party's responsibility can be divided and distributed. Modern legal systems were established on the presupposition that all legal entities are persons. While a robot might someday be considered a person, we are not likely to face this situation any time soon. However, the law has also been designed to deal with several kinds of non-persons, or quasi-persons, and we can look to these for some insights on how we might treat robots that are non-persons, or quasi-persons.

Personhood is a hotly debated concept, and many perspectives in that debate are based in strongly held beliefs from religious faith and philosophical dispositions. Most notably, the case of unborn human fetuses, and the case of severely brain damaged and comatose individuals have led to much debate in the United States over their appropriate legal status and rights. Yet, despite strongly differing perspectives on such issues, the legal systems in pluralistic societies have found ways to deal practically with these and several other border-line cases of personhood.

Minor children (under 18 years of age) are a prime example of quasi-persons. Minors do not enjoy the full rights of personhood that adults do. In particular, they cannot sign contracts or become involved in various sorts of legal arrangements because they do not have the right to do so as

minors. They can become involved in such arrangements only through the actions of their parents or legal guardians. In this sense they are not full legal persons. In another sense, the killing of a child is murder in the same way that the killing of an adult is, and so a child is still a legal person in this sense—and in fact is entitled to many more protections than an adult. Children can thus be considered a type of quasi-person, or legal quasi-agent.

The case of permanently mentally-impaired people can be quite similar to children. Even fully-fledged persons can claim temporary impairments of judgment, and thereby *diminished responsibility* for their actions given certain circumstances, e.g., temporary insanity or being involuntarily drugged. The point is that some aspects of legal agency can apply to entities which fall short of full-fledged personhood and having full responsibility, and it seems reasonable to think that some robots will eventually become a kind of quasi-agent in the view of the law before they achieve full legal personhood.

The concept of personhood is deeply tied to the notion of agency. The law also deals explicitly with agency and, interestingly enough, it does so in order to address cases in which the power of agency is transferred between parties. The law of agency is a highly specialized field that deals mainly with the talent agents of athletes and entertainers, and to some extent insurance, travel, and real estate agents. These agents are empowered by their employers, whom they thereby represent for the purpose of negotiating contracts and making various agreements on their behalf. An individual is bound by the contracts that their agent signs just as if they had signed the contracts themselves, except in cases where one can prove misconduct on the part of the agent. To act as someone's agent is to enact their legal powers from afar, and is in this sense a form of distribution of legal agency.

The possible application to robotics, especially tele-robotics, seems inviting—robots could be seen in many cases as agents acting on the behalf of others. Accordingly, the legal responsibility for the actions of a robot falls on the individual who grants the robot permission to act on their behalf. If it is not already clearly enough implied by the law, it might be advisable to make a law which makes such legal responsibilities explicit. Such a law would need to be carefully crafted, however, to avoid placing too heavy a burden on the owners of robots, preventing the adoption of robots due to risk, and to avoid unfairly protecting manufacturers who might share in the responsibility of misbehaving robots due to poor designs.

5.3 Crime, Punishment, and Personhood

Crime and punishment are central concepts in both law and morality, yet they might seem out of place in a discussion of robot ethics. While we can imagine a humanoid robot of such sophistication

that it is effectively, or indistinguishably, a person, these robots will be easier to find in science fiction than in reality for a long time to come. There are, however, technologically-possible robots that may approach actions that we might consider, at least at first glance, to be criminal. If so, how might the law instruct us to treat such cases?

As stated earlier, criminal law is concerned with punishing wrongdoers, whereas civil law is primarily concerned with compelling wrongdoers to compensate those harmed. There is an important principle underlying this distinction: crimes deserve to be punished, regardless of any compensation to those directly harmed by the crime. Put another way, the harmed party in a crime is the whole of society. Thus, the case is prosecuted by the state or 'the people', and the debt owed by the wrongdoer is owed to the society. While the punishments may take different forms, the point of punishment is traditionally conceived of as being corrective in one or more senses: that the wrongdoer pays their debt to society (retribution); that the wrongdoer is to be reformed so as not to repeat the offense (reform); or that other people in society will be dissuaded from committing a similar wrong (deterrence).

There are two key problems with applying criminal law to robots: (1) criminal actions require a moral agent to perform them, and (2) how is it possible to punish a robot? Moral agency is deeply connected to our concepts of punishment. Moral agency might be defined in various ways, but it ultimately must serve the role of being the subject who is punished. Without moral agency, there can be harm but not guilt. Thus, there is no debt incurred to society unless there is a moral agent to incur it; it is merely an accident or act of nature, but not a crime. Similarly, only a moral agent can be reformed, which implies the development or correction of a moral character; otherwise, it is merely the fixing of a problem. And finally, deterrence only makes sense when moral agents recognize the similarity of their potential choices and actions to those of another moral agent who has been punished for the wrong choices and actions; without this reflexivity of choice by a moral agent, and recognition of similarity between and among moral agents, punishment cannot possibly result in deterrence. There are some interesting ways in which notions of 'training' or 'learning' in artificial intelligence (AI) might be extended to fulfill some aspects of reform and deterrence, however.

In the above, we saw that it is more likely that we will treat robots as quasi-persons long before they achieve full personhood. Lawrence Solum [1992] has given careful consideration to the question of whether an AI might be able to achieve legal personhood, using a thought experiment in which an AI acts as the manager of a trust. He concludes that while personhood is not impossible in principle for an AI to achieve, it is also not clear how we would know that any particular AI has achieved it. The same argument could be applied to robots. Solum imagines a legal Turing test in which it comes down to the determination of a court whether an AI could stand trial as a legal agent in its own right, and not merely a proxy or agent of some other legal entity. He argues that a court would ultimately base its decision on whether the robot in question has moral agency, and whether it is possible to

punish it—could the court fine or imprison an AI that mismanages a trust? In cases of quasi-personhood and diminished responsibility, children and the mentally impaired are usually shielded from punishment as a result of their limited legal status.

There is, however, in the law a relevant case of legal responsibility resting in a non-human, namely the *corporation*. The corporation is a non-human entity that has been effectively granted the legal rights of a person. Corporations can own property, sign contracts, and be held liable for negligence. In certain cases, corporations can even be punished for criminal activities such as fraud, criminal negligence, and causing environmental damage. A crucial aspect of treating corporations as persons depends on the ability to punish them, though this is not nearly so straightforward as it is for human persons. As a 17th century Lord Chancellor of England put it, corporations have "no soul to damn and no body to kick," so how can they be expected to have a moral conscience [Coffee, 1981]?

Of course, corporations exist to make money for themselves or stockholders and as such can be given monetary punishments; and in certain cases, such as anti-trust violations, they can be split apart or dissolved altogether. They cannot be imprisoned, though in criminal cases responsible individuals within the corporation can be prosecuted for their individual actions. As a result of this, and other aspects of corporations being complex socio-technical systems in which there are many stakeholders differently related to the monetary wealth of a corporation, it can be difficult to assign a punishment that achieves retribution, reform, and deterrence while meeting other requirements of fairness, such as proportionality.

Clearly, robots are different in many important respects from corporations. However, there are also many important similarities, and it is no coincidence that John Coffee's [1981] seminal paper on corporate punishment draws heavily on Herbert Simon's [1947] work on organizational behavior and decision making, and in particular how corporate punishment could influence organizational decision making through deterrence. Nonetheless, a great deal of work needs to be done in order to judge just how fruitful this analogy is. While monetary penalties work as punishments for corporations, this is because they target the essential reason for the existence of corporations—to make money. The essential purposes of robots may not be so straightforward, will vary from robot to robot, and may not take a form that can be easily or fairly penalized by a court.

The most obvious difference is that robots *do* have bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment. The various forms of corporal punishment presuppose additional desires and fears central to being human that may not readily apply to robots: pain, freedom of movement, mortality, and so on. Thus, torture, imprisonment, and death are not likely to be effective in achieving retribution, reform, or deterrence in robots. There may be a policy to destroy any robots that do harm; but, as is the case with animals that harm people, it would essentially be a preventative measure to avoid future harms rather than a true punishment.

Whether it might be possible to build in a technological means to enable genuine punishment in robots is an open question.

6. Technology Risk Assessment Framework

Issues related to law and responsibility may be avoided, or better informed, with some forethought to the risks posed by robots. This section will present a framework for evaluating risks arising from of robotic technologies; as this is a preliminary report, we only introduce the primary assessment factors and begin that discovery process, rather than offer comprehensive answers here. The risks we address here are primarily related to harmful but unintended behavior that may arise from robots, though we explore a full range of other risks and issues in section 7 next.

Risk assessment is an interdisciplinary subject, which runs together psychological, ethical, legal, and economic considerations. A major problem in risk assessment is the confusion between popular concepts of risk from robots (the 'subjective risk'), which has largely been made irrational by the various fictional depictions autonomous robots destroying humankind and running amok (as in *Terminator* and *I, Robot*, among many other movies) and the actual objective risk of deploying robots, i.e., what rational basis is there for worry?

First, let us define risk simply in terms of its opposite, safety: **risk** is the probability of harm; and (relative) **safety** is (relative) freedom from risk. Safety in practice is merely relative, not absolute, freedom from harm, because no activity is ever completely risk-free; walking onto one's lawn from inside one's house increases the (however small) risk of death by meteorite strike. Hence, risk and safety are two sides of the usual human attempt to reduce the probability of harm to oneself and others. War is a strange human activity not least because it reverses this tendency; in war, one wishes to increase the probability of harm to one's enemies. But the Laws of War make clear that not all ways of increasing risk for one's enemy are morally legitimate; and some ways of increasing risk for one's own side may be morally legitimate and even morally required. These facts considerably complicate the ethics of risk assessment for military robots.

6.1 Acceptable-Risk Factor: Consent

To begin, the major factors in determining 'acceptable risk' in robotics, including military robots, will include (but are not limited to):

> *Consent*: Is the risk voluntarily endured, or not? For instance, secondhand smoke is more objectionable than firsthand, because the passive smoker did not consent to the risk even if

the objective risk is smaller. Will those who are at risk from work with robots reasonably give consent? When (if ever) would it be appropriate to deploy or use robots without consent of those affected?

Morality requires the possibility of consent; to be autonomous is at a minimum to have the capacity to either give or withhold consent to some action. On this basis, Robert Sparrow has a critique of the very possibility of morally deploying autonomously-functioning military robots [Sparrow 2007]: his contention is that such robots can never be morally deployed, because no one—neither the programmer, nor the commanding officer, nor the robot itself—can be held responsible if it commits war crimes or otherwise acts immorally. No one can reasonably be said to give morally responsible consent to the action an autonomous robot performs; so no one is responsible for the risk such autonomous robots pose, and thus it is immoral to use them.

One can imagine a response to Sparrow as follows: We find it morally permissible for military parents to raise their child as destined for the military, to indoctrinate them as a soldier from infancy, and to place those expectations on them in their earliest training. Once they become autonomous adults, it is expected that they will volunteer for service—but they remain autonomous, and it is possible (however psychologically unlikely) that they will choose a different path. If it is morally permitted to raise human children with such expectations, and to accept the children so indoctrinated into voluntary military service, why it would be wrong to likewise train an autonomous robot and place it into active duty?

But some may object as follows: A human child will develop free will, and the above analogy fails given a robot's lack of true Kantian autonomy. That is, the robot could never have the sense of self or the libertarian free will of humanity; they have merely instrumental (means-ends, goal-oriented) rationality. Humans are not robots and have a different kind of autonomy than robots ever could.

6.1.1 A Solution to Sparrow's Problem: Robots as Slaves

Leaving aside those who think humans really are merely complex robots [e.g., Dennett, 1995], there is a simpler solution to Sparrow's objection to the 'in-principle' immorality of deploying autonomous (in the sense of self-regulating) robots. For all military robots, including those with this minimal self-regulating level of autonomy, we normally assume what the literature terms a '*slave morality*', i.e., they have no ends of their own, but their goals are all in service of the goals of someone else—in this case, the military and, more specifically, whoever commands them and gives their orders. Robots cannot create their own laws or final goals; they are not ultimately makers of a *self*, but followers of 'life goals' others have imposed, and their own freedom only comes in the mere means they choose to realize those ends.

Such military robots, whatever their other decision-making capabilities, thus lack full Kantian autonomy, and so cannot be held responsible for their actions under traditional deontological, natural law, or virtue ethics theories. Inasmuch as *jus in bello* restrictions most plausibly depend on one of those approaches, robot risk and responsibility as a function of consent thus becomes a non-issue. This realization helps to rebut the central contention of Sparrow's critique of autonomously functioning military robots.

Again, Sparrow's contention is that such robots can never be morally deployed, because no one—neither the programmer, the commanding officer, nor the robot itself— can be held responsible if it commits war crimes or otherwise acts immorally, because no one *Self* is in control of what happens. But as long as a slave morality is built in to these otherwise autonomous robots, the basis for Sparrow's objection is undermined: the robot cannot be blamed, for it really is 'merely following orders', subject to the limitations of its programming. It could not become a *morally* autonomous 'law unto itself' and serve its own ends; hence, it cannot be held morally responsible for its actions.

Of course, one immediate objection could point to Nazi soldiers who committed war crimes and pleaded that they were only following orders; and we can imagine back in the day that George Washington's slaves might have been held responsible for following immoral orders. But there is a crucial difference between a human soldier or even a human slave and a robot programmed with a slave morality: the human person, whether a soldier or a slave, is presumed to have the ability to disobey orders, even if the punishment for doing so would be harsh. But a properly programmed robot with a 'slave morality' literally could not disobey orders intentionally—it would do so only by mistake. And in ethics, as long as someone is not free (in whatever is the relevant moral sense of 'free') to disobey their orders, they cannot be blamed. For robots, unlike humans, that can be a matter of correct programming.

6.1.2 Machine Learning and Consent?

A further possible objection, then: What of the unpredictability arising from 'machine learning'? Could that enable robot consent and hence robot responsibility? Perhaps Sparrow's concern is not so much about robots that strictly follow their programming, but more about robots programmed to learn and create their own framework for making decisions based on what was learned, as previously discussed in section 3 of this report.

But this concern, while legitimate, does not succeed at moving responsibility from the commanding officer to the suitably programmed robot. The relevant moral sense of 'free' that moral responsibility entails seems to involve Kantian autonomy, i.e., the freedom to choose one's own life goals for oneself. What a slave morality amounts to is not the absence of a freedom of means, i.e., of choosing, from among alternatives, the best means to attaining one's ends and learning better

means to those ends; instead, slave morality entails an inability to choose one's own ends, i.e., a lack of true Kantian autonomy. That is compatible with machine learning; the machine will learn the best means for obtaining its preprogrammed goals, but it will *not* be able to overwrite those goals (in the military context, that means it will not be able to overwrite the LOW and ROE). In other words, part of its program will be subject to self-revision (machine learning), but another part (that establishes its unchangeable goals, i.e., the LOW and ROE) will not be.

However, the person who gave the orders to the robot is a Self and has a choice as to whether to create and/or deploy this robot with a limited freedom towards achieving the commander's ends. Hence, the robot is a *tool,* and ethics and military law both accept that one is responsible for one's choice of tools for accomplishing one's ends. So, as the military officially desires, the commanding officer is rightly to blame for any crimes his robotic *slave* chooses to commit, in virtue of his choice to deploy the slave. An exception would be one due to deliberate misprogramming or faulty manufacture that was unknown by the commanding officer, in which case the programmer or manufacturer would be responsible (see the previous discussion about liability law in section 5). Otherwise, the only one voluntarily consenting to the risk—and therefore the only party that can be held responsible—is the chain of command and, specifically, the commanding officer.

Thus, to create a robot capable of the type of consent required for moral responsibility in risk-taking, we must create a Kantian-autonomous robot—but even if that were possible, creating such a robot cannot possibly yet be justified from an 'acceptable risk' ethical perspective. Relatedly, a crucial risk to be avoided in making the deployment of robots morally acceptable is at all costs to avoid the possibility of *rampancy*, i.e., an AI overwriting its own programming, at least as regards the most fundamental aspects of its goals, such as the LOW and ROE. Such a robot would have the potential to leave behind its imposed slave morality and become autonomous in the Kantian sense: the programmer of its own self and own goals, or the maker of its own destiny. Not only would such robots pose incredible risks to humans in the possibility of rampancy, but they would also be undesirable from a military ethics and responsibility perspective: they would then move moral responsibility from the commanding officer to the robot itself. But the refusal (and current inability) to create a Kantian-autonomous robot solves Sparrow's dilemma. So for the foreseeable future, we solve both the problems of risk and responsibility by requiring a slave morality.

6.2 Acceptable-Risk Factor: Informed Consent

So, given a robotic slave morality, the only one consenting to the risk of military robotic malfeasance is the military command; but this still leaves unanswered the query as to whether the risk (of malfunction or other error) to the *unintended* targets—the noncombatants—is morally permissible. After all, the noncombatants clearly did not consent to the deployment of the robot. Perhaps self-

regulating military robots will be immoral to deploy because of the risk they pose to noncombatants. To assess this possibility, we need a further investigation into risk assessment, especially as regards involuntary or non-voluntary risks. In order to do so, we first examine another issue involving consent: Does the morality of consent require adequate knowledge of what is being consented to?

> *Informed consent*: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from robots know they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the 'known unknowns') be handled?

The risk for the military in using autonomous robots and for the civilian population can thus be detailed more precisely: Is the military command obligated to inform the civilian noncombatant population that self-regulating robots are being deployed, and the nature of the risk they pose? Likewise, is the military command similarly obligated to inform its own soldiers (or the enemy's soldiers) that self-regulating robots are being deployed, and the nature of the risk they pose? The last part is the simplest: Under the Laws of War, enemy combatants have no general right to know the nature of the weapons being used against them. Surprise is well understood as a legitimate tactic in war.

Similarly, while the military obviously has a self-interest in making its own soldiers safe from the risk of malfunctioning robots, soldiers in general have no *right* to safety from the military's own weapons. From insufficiently-armored personnel vehicles to friendly fire, military personnel know they are at risk from their own side as well as the enemy. The moral as well as practical requirement for the military command is to minimize that risk to one's own side; a large part of the push for the deployment of self-regulating military robots is precisely the hope that they can reduce such risks to human soldiers on one's own side, not increase them.

Finally, just as the enemy combatants have no *right to know* the exact nature of the weapons and risks arrayed against them, so too the Laws of War have denied civilian noncombatants any such right to know their level of risk. Inasmuch as targeting them is immoral, their risk is one of 'collateral damage'; and *jus in bello* restrictions demand only that military weapons and tactics attempt to minimize such collateral damage—they do not require the civilians to have a precise knowledge of the risks. (Indeed, an attempt to explain the nature and severity of the risks of collateral damage to the enemy's civilian population could well be seen as an act of terrorism!) Further, as Arkin maintains, it is entirely possible that deploying military robots will not only reduce the risk of harm to one's own troops, but if suitably programmed, it conceivably could reduce the risk of collateral

damage [Arkin, 2007, p.57]. Hence, the morality of the risk of deploying military robots does not turn on issues of informed consent.

6.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to military robots, we may continue to focus on the affected population as another factor in determining acceptable risk:

> *Affected population*: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? In military terms, civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary. Will the military use of robots pose the risk of any special harms to noncombatants?

As Arkin maintains, the issues here depend on how the robots are programmed and how reliable they are [Arkin, 2007, pp.57-60]. Assuming the LOW and ROE are suitably programmable, Arkin plausibly argues that robots would decrease the risk to noncombatants; this also assumes sufficient and suitably realistic pre-deployment testing to alleviate *first generation* problems, i.e., while it is morally unjustifiable to deploy military robots before we have any idea of their risk to noncombatants, we may paradoxically need to use the first deaths to determine the level of risk (see below and section 7 for more on this).

What this aspect of risk assessment makes clear is that the standards of safety with respect to noncombatants are likely to be quite high; until and unless military robots are capable of having a risk of collateral damage on parity with (or better than) human soldiers, there will be serious moral qualms in deploying them under generally accepted *jus in bello* restrictions. Robotic weapons that attack indiscriminately or disproportionately—similar in effect to landmines as well as nuclear, biological, and chemical weapons—are hence immoral to deploy. Whether or not robotic weaponry will soon be able to technologically meet the moral imperative to minimize collateral damage is one of the foremost issues in the ethics of autonomous military robots.

6.4 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment: seriousness and probability, or how bad would the harm be, and how likely is it to happen?

Seriousness: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive. What (if any) serious risks from robots are acceptable—and to whom: soldiers, noncombatants, the environment, or the robots themselves?

Probability: This is often conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from the robots? How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?

The *jus in bello* tradition of emphasizing the requirements of discrimination and proportionality in military weaponry provide a guidepost here. The *seriousness* of risk can be given at least a rough operational definition in terms of the already understood concept of *proportional response*; it is already accepted in combat that soldiers legitimately run the risk of murder (but not, e.g., torture) by the enemy, and hence there should be no moral qualms in principle about lethal military robots—whereas an automated torture device would rightly be morally condemned. But it is also understood that morally legitimate warfare does not seek the superfluous deaths of the enemy, and so the seriousness of the risk that robots pose should be adequate to the military objective, *but no greater*. Again, whether or not robotic weaponry will soon be able to surmount the technical challenge of this moral imperative (at least as well as human soldiers) remains unknown. Likewise, what has been said above about risks to noncombatants pertains to the seriousness of their risk: unless military robots plausibly pose no more serious a risk to them than the ordinary human seriousness of collateral damage, deploying the robots will be immoral, under *jus in bello*.

Yet more complex may be the issue of the *probability* of harm. In general, weapons and tactics that increase the probability of harm to the enemy are considered good; a weapon that guarantees the death of the enemy would be considered desirable– assuming it does not also guarantee harm to noncombatants. But creating weapons that are both highly lethal and highly discriminating has proven difficult; it is entirely possible that robots will prove a breakthrough here and be amply morally justifiable. But the issue of certainty is a key here, especially as regards the first-generation problem; it seems clear that extensive pre-deployment testing will be required to ensure military robots only raise the probability of harm to the enemy and pose only an acceptable threat to noncombatants.

6.5 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In all social theorizing, concepts have a certain degree of fluidity, dependent upon how those in power determine their meaning. The concept of risk, which includes psychological, legal, and economic considerations as well as ethical ones, is certainly no different. Hence, the concept of an acceptable risk—or an unacceptable one—is at least in part socially constructed. (And so proposing a survey of what Americans have believed and defended about acceptable risk may help answer the question of what risks are acceptable.)

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable:

> *Good faith subjective standard*: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can soldiers in the battlefield be trusted to make wise choices about (un)acceptable risk? This seems incompatible with the moral deployment of autonomous, non-tele-operated (no 'human in the loop') robots, for reasons having to do with the inevitability of robot mistakes—see below. The problem of nonvoluntary risk borne by civilian noncombatants makes this standard impossible to defend, as does the idiosyncrasies of human risk aversion.
>
> *The reasonable-person standard*: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute military regulations or some other basis for what a 'reasonable person' would think for the difficult-to-foresee vagaries of conditions in the field and the subjective judgment of soldiers? Or what kind of judgment would we expect an autonomous robot to have—would we trust it to accurately determine and act upon the assessed risk? If not, then autonomous robots could never be deployed without tele-operators—that is, without a human in the loop. Even a 'kill switch' that enabled autonomous operation until a remote surveillance operator determined something had gone wrong (and could disable the robot, or at least its autonomous functioning) would only come into effect after something had already gone wrong, i.e., the first-generation problem.
>
> *Objective standard*: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there is still the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

It seems clear enough that as regards the military use of autonomous robots, only the last standard has any plausibility. It is also the standard most often defended in law and practice; but it does have that serious first-generation problem. Fortunately, there is a solution. Simply put, to use the objective standard of risk assessment, we then have an ethical obligation for extended testing of self-regulating, autonomous robots in artificial and human-free environments before risking robot-human interaction. This testing must be thorough, extensive, realistic, variegated, and come in stages, so that full deployment with possible or actual civilian contact comes only at the end of a long training regimen and safety inspection. Such extended testing could never guarantee that autonomous robots would not make horrible mistakes in the confusing, hard-to-foresee, and data-intensive fog of war; there is no possibility of taking something without the possibility of *mis*-taking. But such testing could give us effective rational confidence that such mistakes would be less than those made by humans in similar situations. From a risk-reward perspective, it seems clearly acceptable to deploy autonomous robots as soon as such extensive testing indicated their mistakes were, on average, no worse than (or better than) the typical human soldier.

This solution to the first-generation problem indicates the obvious way forward. We cannot trust humans to determine risks for autonomous robots, not least because we are often psychologically, emotionally, and cognitively ill-equipped to accurately understand and estimate the risk. As robots grow in their lethality, speed, and autonomy, this problem will only become more acute. One of the few near-certainties in the development of military robots is that keeping a human in the decision-making loop is going to seriously degrade battle efficiency soon—and it may likely also degrade risk assessment. Military robots, for better or worse, may soon have better capabilities to judge real-time risks than their teleoperator sitting thousands of miles away. With sufficient research and pre-deployment testing, the objective features of those risks and a decision algorithm for their assessment can be programmed that gives such robots human-equivalent or better risk assessment capabilities. At this stage, we need to make it a moral imperative that such capacities are so programmed before these robots are actually deployed. Such an approach should resolve the worries about safety and dependability concerns prominent in the literature [e.g., Sharkey, 2007a; Van der Loos, 2007].

Assuming we combine this resolve to engage in serious, realistic, and extensive pre-deployment simulation testing with a requirement for a 'learning curve' in which robots must pass a series of increasingly difficult tests before deployment, most of the main risk concerns should be alleviated. Such extensive testing will further resolve issues about the unpredictability of the behavior of deployed robots and their ability to manage complex, hostile environments. If they prove unequal to safe deployment in testing, it may simply be immoral to deploy them.

6.6 Other Risks

There remains a perpetual risk concerning security issues for autonomous robots, although the issues here are common to many aspects of technological culture and are hardly unique to autonomous robots. For example, how susceptible would a military robot be to hacking or reprogramming after capture? If it could be reprogrammed in any of the ways deemed prohibited above, that would be a serious risk and reason to avoid deployment. There are related risks that are specific to military autonomous robots: for instance, are the Rules of Engagement and the Geneva Convention actually reducible to algorithms (or, more plausibly, algorithms plus machine learning)? If so, is that enough to ensure ethical conduct in robots?

And finally, some have raised risks of a more abstract sort, indicating the rise of such autonomous robots creates risks that go beyond specific harms to societal and cultural impacts. For instance, is there a risk of (perhaps fatally?) affronting human dignity or cherished traditions (religious, cultural, or otherwise) in allowing the existence of robots that make ethical decisions? Do we 'cross a threshold' in abrogating this level of responsibility to machines, in a way that will inevitably lead to some catastrophic outcome? Without more detail and reason for worry, such worries as this appear to commit the 'slippery slope' fallacy. But there is worry that as robots become 'quasi-persons' [Asaro, 2007], even under a 'slave morality', there will be pressure to eventually make them into full-fledged Kantian-autonomous persons, with all the risks that entails.

What seems certain is that the rise of autonomous robots, if mishandled, will cause popular shock and cultural upheaval, especially if they are introduced suddenly and/or have some disastrous safety failures early on. That is all the more reason that a lengthy period of rigorous testing and gradual rollout (crawl-walk-run approach) is a moral minimum for the ethical deployment of autonomous robots, especially by the military. Further, this points to the early, prior need to identify the full range of possible ethical, technological, and societal issues in robot ethics—as we will discuss in the next section—in order to ensure that a technology risk assessment accounts for these concerns.

7. Robot Ethics: The Issues

From the preceding sections, it should be clear that there are myriad issues in risk and ethics related to autonomous military robotics. In this section, we will pull together and organize these various strands, as well as raise additional ones to provide a single, full view of the challenges facing the field.[6] These challenges are organized in thematic sub-groups: legal, just war, technical, robot-human, societal, and other and future challenges.

This is not meant to be an exhaustive list, as other issues certainly will emerge as the technology develops and field use broadens.[7] The value of this section again is to help anticipate the challenges facing the development and deployment of autonomous military robots, in order to proactively address them in both the design or application phases. Further, they may help to inform ethical and risk issues related to non-military robots, given the close historical relationship between defense technologies and consumer or public technologies, such as the evolution of ARPANET into the Internet.

7.1 Legal Challenges

1. *Unclear responsibility*. To whom would we assign blame—and punishment—for improper conduct and unauthorized harms caused by an autonomous robot (whether by error or intentional): the designers, robot manufacturer, procurement officer, robot controller/supervisor, field commander, President of the United States...or the robot itself? [Asaro, 2007; Sparrow, 2007; Sharkey, 2008a] We have started an inquiry into this critical issue in section 5: The law offers several precedents that a robotics case might follow, but given the range of specific circumstances that would influence a legal decision as well as evolving technology, more work will be needed to clarify the law for a clear framework in matters of responsibility.

[6] We thank and credit Ron Arkin for his discussions on many of these issues presented here.
[7] As an example of an unexpected policy change, when German forces during World War II recognized the impracticality of using naval submarines to rescue crews of sinking enemy ships—given limited space inside the submarine as well as exposure to radar and attacks when they surface—they issued the *Laconia* Order in 1942, based on military necessity, that released submarines from a long-standing moral obligation for sea vessels to rescue survivors; other nations soon followed suit to effectively eliminate the military convention altogether [Walzer, 1977, pp. 147-151].

In a military system, it may be possible to simply *stipulate* a chain of responsibility, e.g., the commanding officer is ultimately responsible. But this may oversimplify matters, e.g., inadequate testing allowed a design problem to slip by and caused the improper robotic behavior, in which case perhaps a procurement officer or the manufacturer ought to be responsible. The situation becomes much more complex and interesting with robots that have greater degrees of autonomy, which may make it appropriate to treat them as quasi-persons, if not full moral agents some point in the future. We note that Kurzweil forecasts that, by the year 2029, "[m]achines will claim to be conscious and these claims will be largely accepted" [Kurzweil, 1999].

2. *Refusing an order*. A conflict may arise in the following situation, among others: A commander orders a robot to attack a house that is known to harbor insurgents, but the robot—being equipped with sensors to 'see' through walls—detects many children inside and, given its programmed instruction (based on the ROE) to minimize civilian casualties, refuses the order. How ought the situation proceed: should we defer to the robot who may have better situational awareness, or the officer who (as far as she or he knows) issues a legitimate command? This dilemma also relates back to the question of responsibility: if the robot refuses an order, then who would be responsible for the events that ensue? Following legitimate orders is clearly an essential tenet for military organizations to function, but if we permit robots to refuse an order, this may expand the circumstances in which human soldiers may refuse an order as well (for better or worse).

3. *Consent by soldiers to risks*. In October 2007, a semi-autonomous robotic cannon deployed by the South African army malfunctioned, killing nine 'friendly' soldiers and wounding 14 others [Shachtman, 2007]. It would be naive to think such accidents will not happen again. In these cases, should soldiers be informed that an unusual or new risk exists, e.g., when they are handling or working with other dangerous items, such as explosives or even anthrax? Does consent to risk matter anyway, if soldiers generally lack the right to refuse a work order? We discussed the notion of consent and informed in the previous section.

7.2 Just-War Challenges

1. *Attack decisions*. It may be important for the above issue of responsibility to decide who, or what, makes the decision for a robot to strike. Some situations may develop so quickly and require such rapid information processing that we would want to entrust our robots and systems to make critical decisions. But the LOW and ROE generally demand there to be human 'eyes on target', either in-person or electronically and presumably in real time. (This is another reason why there is a general ban on landmines: without eyes on target, we do not know who is harmed

by the ordnance and therefore have not fulfilled our responsibility to discriminate combatants from non-combatants.) If human soldiers must monitor the actions of each robot as they occur, this may limit the effectiveness for which the robot was designed in the first place: robots may be deployed precisely because they can act more quickly, and with better information, than humans can.

However, some military robots—such as the Navy's Phalanx CIWS—seem to already and completely operate autonomously, i.e., they make attack decisions without human eyes on target or approval. This raises the question of how strictly we should take the 'eyes on target' requirement. One plausible argument for stretching that requirement is that the Phalanx CIWS operates as a last line of defense against imminent threats, e.g., incoming missiles in the dark of the night, so the benefits more clearly outweigh the risks in such a case. Another argument perhaps would be that 'eyes on target' need not be *human* eyes, whether directly or monitoring the images captured by a remote camera; that is, a human does not necessarily need to directly confirm a target or authorize a strike. A robot's target-identification module—assuming it has been sufficiently tested for accuracy—programmed by engineers is arguably a proxy for human eyes. At least this gives the system some reasonable ability to discriminate among targets, in contrast to a landmine, for instance. A requirement for 100% accuracy in target identification may be overly burdensome, since that is not a bar we can meet with human soldiers.

2. *Lower barriers for war*. As raised in section 4, does the use of advanced weaponry such as autonomous robotics make it easier for one nation to engage in war or adopt aggressive foreign (and domestic) policies that might provoke other nations? If so, is this a violation of *jus ad bellum*? [Asaro, 2008; Kahn, 2002] It may be true that new strategies, tactics, and technologies make armed conflict an easier path to choose for a nation, if they reduce risks to our side. Yet while it seems obvious that we should want to reduce US casualties, there is something sensible about the need for some terrible cost to war as a deterrent against entering war in the first place. This is the basis for just-war theory, that war ought to be the very last resort given its horrific costs [Walzer, 1977].

   But the considered objection—that advanced robotics immorally lowers barriers for war—hides a logical implication that we should not do anything that makes armed conflict more palatable: we should not attempt to reduce friendly casualties, or improve battlefield medicine, or conduct any more research that would make victory more likely and quicker. Taken to the extreme, the objection seems to imply that we should *raise* barriers to war, to make fighting as brutal as possible (e.g., using primitive weapons without armor) so that we would never engage in it unless it were truly the last resort. Such a position appears counterintuitive at best and dangerously foolish at worst, particularly if we expect that other nations would not readily adopt a policy of relinquishment, which would put the US at a competitive disadvantage.

3. *Imprecision in LOW and ROE*. Asimov's Laws appear to be as simple as programmable rules can be for autonomous robots, yet they yielded surprising, unintended implications in his stories [e.g., Asimov, 1950]. Likewise, we may understand each rule of engagement and believe them to be sensible, but are they truly consistent with one another and sufficiently clear—which appears to be a requirement in order for them to be programmable? Much more complex than Asimov's Laws, the LOW and ROE leave much room for contradictory or vague imperatives, which may result in undesired and unexpected behavior in robots.

   For instance, the ROE to minimize collateral damage is vague: is the rule that we should not attack a position if civilian deaths are expected to be greater than—or even half of— combatant deaths? Are we permitted to kill one (high-ranking) combatant, even if it involves the death of five civilians—or $10M in unnecessary damage? A robot may need specific numbers to know exactly where this line is drawn, in order to comply with the ROE. Unfortunately, this is not an area that has been precisely quantified nor easily lends itself for such a determination.

7.3 Technical Challenges

1. *Discriminating among targets.* Some experts contend that it is simply too difficult to design a machine that can distinguish between a combatant and a non-combatant, particularly as insurgents pose as civilians, as required for the LOW and ROE [e.g., Sharkey, 2008a; Sparrow, 2007; Canning et al., 2004]. Further, robots would need to discriminate between active combatants and wounded ones who are unable to fight or have surrendered. Admittedly, this is a complex technical task, but we need to be clear on how accurate this discrimination needs to be. That is, discrimination among targets is also a difficult, error-prone task for human soldiers, so ought we hold machines to a higher standard than we have yet to achieve ourselves, at least in the near term?

   Consider the following: A robot enters a building known to harbor terrorists, but at the same time an innocent girl is running toward the robot (unintentionally) in chasing after a ball that happens to be rolling in the direction of the robot. Would the robot know to stand down and not attack the child? If the robot were to attack, of course that would cause outrage from opposing forces and even our own media and public; but this scenario could likely be the same as with a human soldier, adrenaline running high, who may misidentify the charging target as well. It seems that in such a situation, a robot may be less likely to attack the child, since the robot is not prone to overreact from the influence of emotions and fear, which afflict human soldiers. But in any event, if a robot would likely not perform worse than a human soldier, perhaps this is good enough for the moment until the technical ability to discriminate among

targets improves. Some critics, however, may still insist on perfect discrimination or at least far better than humans are capable of, though it is unclear why we should hold robots to such a high standard before such a technology exists (unless their point is to not use robots at all until we have perfected them, which is also a contentious position).

One proposed 'workaround' solution is to permit robots to target only weapons, including any hostile robots, rather than the human soldiers wielding such weapons [Canning, 2008]. Thus if an enemy combatant fails to relinquish his weapon in the presence of a robot, then he significantly increases his risk of being unintentionally harmed as the robot proceeds to disable the weapon. However, while this seems reasonable in principle, other experts continue to point to the technical challenge of discrimination: given current and foreseeable limitations in AI, a robot still may not be able to reliably target only a weapon and not the person, nor even reliably identify weapons from non-weapons, e.g., a child pointing her ice cream cone at an urban patrol robot [Sharkey, 2007b]. If this is true, then the considered solution merely postpones the discrimination problem, though it does create an extra layer of protection against inappropriate harm to humans; so the solution merits further consideration.

Another possible solution, which avoids the above programming issues, may be to simply operate combat robots only in regions of heavy fighting, teeming with valid targets [Sharkey, 2008b]. In these zones—sometimes called 'kill boxes' or 'engagement regions'—the Rules of Engagement are loosened, and non-combatants can be reasonably presumed to have fled, thus obviating the issue of discriminating among targets (and assuming none of our own troops is in the kill box or at least can be easily identified, e.g., by wearing some sensor). Using combat robots, at least initially, in only such zones might help to solve the first-generation problem described below, providing a training ground of sorts to test and perfect the machines. Or even in regions without heavy fighting but in need of tight security, e.g., guarding perimeters, armed sentry robots could operate in those designated zones, as long as it is clear to everyone that trespassers will be presumed to be enemy combatants and sufficient safeguards or deterrents to entry are in place to prevent, say, an accidental trespass by a child. (The risk of harm to a non-combatant here seems to be the same as with using guard dogs to protect property today.)

2. *First-generation problem.* We previously mentioned that it would be naive to believe that another accident with military robots will not happen again. As with any other technologies, errors or bugs will inevitably exist, which can be corrected in the next generation of the technology. With Internet technologies, for instance, first-generation mistakes are not too serious and can be fixed with software patches or updates. But with military robotics, the stakes are much higher, since human lives may be lost as a result of programming or other errors. So it seems that the prudent or morally correct course of action is to rigorously test the robot before deploying it, as discussed in section 6.

However, testing already occurs with today's robots, yet it is still difficult if not impossible to certify any given robot as error-free, given that (a) testing environments may be substantially different than more complex, unstructured, and dynamic battlefield conditions in which we cannot anticipate all possible contingencies; and (b) the computer program used in the robot's on-board computer (its 'brain') may consist of millions of lines of code.

Beta-testing of a program (testing prior to the official product launch, whether related to robotics, business applications, etc.) is conducted today, yet new errors are routinely found in software by actual users even after its official product launch. It is simply not possible to run a complex piece of software through all possible uses in a testing phase; surprises may occur during its actual use. Likewise, it is not reasonable to expect that testing of robots will catch any and all flaws; the robots may behave in unexpected and unintended ways during actual field use. Again, the stakes are high with deploying robots, since any error could be fatal. This makes the first-generation problem, as well as ongoing safety and dependability, an especially sensitive issue [e.g., Van der Loos, 2007].

3. *Robots running amok*. As depicted in science-fiction novels and movies, some imagine the possibility that robots might break free from their human programming through methods as: their own learning, or creating other robots without such constraints (self-replicating and self-revising), or malfunction, or programming error, or even intentional hacking [e.g., Joy, 2000]. In these scenarios, because robots are built to be durable and even with attack capabilities, they would be extremely difficult to defeat—which is the point of using robots as force multipliers. Some of these scenarios are more likely than others: we wouldn't see the ability of robots to fully manufacture other robots or to radically evolve their intelligence and escape any programmed morality for quite some time. But other scenarios, such as hacking, seem to be near-term possibilities, especially if robots are not given strong self-defense capabilities (see below).

   That robots might run amok is an enhanced version of the worry that enemies might use our own creations against us, but it also introduces a new element in that previous weapon systems still need a human operator which is a point of vulnerability, i.e., a 'soft underbelly' of the system. Autonomous robots would be designed to operate without human control. What precautions can be taken to prevent one from being captured and reverse-engineered or reprogrammed to attack our own forces? If we design a 'kill switch' that can automatically shut off a robot, this may present a key vulnerability that can be exploited by the enemy.

4. *Unauthorized overrides*. This concern is similar to that with nuclear weapons: that a rogue officer may be enough to take control of these terrible weapons and unleash them without

authorization or otherwise override their programming to commit some unlawful action. This is a persistent worry with any new, devastating technology and is a multi-faceted challenge: it is a human problem (to develop ethical, competent officers), an organizational problem (to provide procedural safeguards), and technical problem (to provide systemic safeguards). So there does not yet appear to be anything unique about this worry that should hinder the development or deployment of advanced robotics, to the extent that the concern does not impact the development of other technologies. But it nevertheless is a concern that needs to be considered in the design and deployment phases.

5. *Competing ethical frameworks*. If we seek to build an ethical framework for action in robots, it is not clear which ethical theory to use as our model [e.g., Anderson and Anderson, 2007]. In section 3, we have argued for a hybrid approach related to virtue ethics, as the theory that seems to lead to the fewest unintuitive results, but any sophisticated theory seems to be vulnerable to inconsistencies and competing directives (especially if a three- or four-rule system as simple as Asimov's cannot work perfectly). This concern is related to the first technical challenge described here, that it is too difficult to embed these behavioral rules or programming into a machine. But we should recall our stated mission here: our initial goal ought *not* be to create a perfectly ethical robot, only one that acts more ethically than humans—and sadly this may be a low hurdle to clear.

6. *Coordinated attacks*. Generally, it is better to have more data than less when making decisions, particularly one as weighty as a military strike decision. Robots can be designed to easily network with other robots and systems; but this may complicate matters for robot engineers as well as commanders. We may need to establish a chain of command within robots when they operate as a team, as well as ensure coordination of their actions. The risk here is that as complexity of any system increases, the more opportunities exist for errors to be introduced, and again mistakes by military robots may be fatal.

7.4 Human-Robot Challenges

1. *Effect on squad cohesion*. As a 'band of brothers', there understandably needs to be strong trust and support among soldiers, just as there is among police officers, firefighters, and so on. But sometimes this sense of camaraderie can be overdeveloped to the extent that one team member becomes complicit in or deliberately assists in covering up an illegal or inappropriate action of another team member. We have discussed the benefits of military robots with respect to behavior that is more ethical than currently exhibited by human soldiers. But robots will also likely be equipped with video cameras and other such sensors to record and report actions on the battlefield. This could negatively impact the cohesion among team or squad members by

eroding trust with the robot as well as among fellow soldiers who then may or may not support each other as much anymore, knowing that they are being watched. Of course, soldiers and other professionals should not be giving each other unlawful 'support' anyway; but there may be situations in which a soldier is unclear about or unaware of motivations, orders, or other relevant details and err on the side of caution, i.e., not providing support even when it is justified and needed.

2. *Self-defense*. Asimov's Laws permitted robots to defend themselves where that action did not conflict with higher duties, i.e., harm humans (or humanity) or conflict with a human-issued order. But Arkin suggests that military robots can be more conservative in their actions, i.e., hold their fire, because they do not have a natural instinct of self-preservation and may be programmed without such [Arkin, 2007]. But how practical is it, at least economically speaking, to not give robots—which may range from $100,000 to millions of dollars in cost—the ability to defend themselves? If a person, say, a US civilian, threatens to destroy a robot, shouldn't it have the ability to protect itself, our very expensive taxpayer-funded investment?

   Further, self-defense capabilities may be important for the robot to elude capture and hacking, as previously discussed. Robots may be easily trapped and recovered fully intact, unlike tanks and aircraft, for instance, which usually sustain much if not total damage in order to capture it. These considerations are in tension with using robots for a more ethical prosecution of war, since a predilection to hold their fire would be a major safeguard against accidental fatalities, e.g., mistakenly opening fire on non-combatants; therefore, a tradeoff or compromise among these goals—to have a more ethical robot and to protect the robot from damage and capture—may be needed.

3. *Winning hearts and minds*. Just-war theory, specifically *jus post bellum*, requires that we fight a war in such a manner that it leaves the door open for lasting peace after the conflict [Orend, 2002]. That is, as history has shown, we should not brutalize an enemy, since that would leave ill-feelings to linger even after the fighting has stopped, which makes peaceful reconciliation most difficult to achieve. Robots do not necessarily represent an immoral or overly brutal way of waging war, but as they are needed for urban operations, such as patrolling dangerous streets to enforcing a curfew or securing an area, the local population may be less likely to trust and build good-will relationships with the occupying force [Sharkey, 2008a]. Winning hearts and minds is likely to require diplomacy and human relationships that machines would not be capable of delivering at the present time, as we previously discussed in section 4.

4. *'Comfort' robots.* Ethicists are already talking about the impact of robots as lovers or surrogate relationship partners [Levy, 2007]. This does not seem so unthinkable, considering that some people already have 'relationships' with increasingly-realistic sex dolls, so robotics appear to be

a natural next step in that industry; indeed, people today engage in sexual activities online, i.e., without a partner physically present.

In previous wars, women have been taken by the military to provide 'comfort' to soldiers or, in other words, forced into sexual slavery or prostitution. In World War II, women were most infamously used by the Japanese Imperial Army to satiate the pent-up carnal desires of its soldiers, ostensibly to prevent possible riots and discontent among the ranks; Nazi Germany reportedly also used women to stock their 'joy divisions' at labor or concentration camps. And instances of rape have been reported—and continue today—in armed conflicts from Africa to the Americas to Asia.

Robots, then, may be able to serve the same function of providing 'comfort' to the troops in a much more humane way, i.e., without the exploitation of women and prisoners of war. However, it is unclear that this function is truly needed (to the extent that most militaries today do not employ military prostitutes and seem to be operating adequately) or can overcome existing public inhibitions or attitudes on what is mostly a taboo subject of both sex in the military and sex with non-human objects.

7.5 Societal Challenges

1. *Counter-tactics in asymmetric war*. As discussed in the previous issue of lowering barriers to war or making war more risk-free, robots would help make US military actions more effective and efficient, which is exactly the point of deploying those machines. Presumably, the more autonomous a robot is, the more lethal it can be (given requirements to discriminate among targets and so on). This translates to quicker, more decisive victories for us; but for the other side, this means swifter and perhaps more demoralizing defeats. We can reasonably expect that a consequence of increasing the asymmetry of warfare in our favor will cause opposing forces to engage in even more unconventional strategies and tactics, beyond 'terrorist' acts today as necessitated by an overwhelming superiority of US troop numbers and technologies [e.g., Kahn, 2002]; few nations could hope to successfully wage war with the US by using the same methods we use.

   This not only involves how wars and conflicts are fought, but also exposes our military as well as public to new forms of attack which may radically change our society, as the events of 9/11 have already. For instance, more desperate enemies may resort to more desperate measures, from intensifying efforts to acquire nuclear or biochemical weapons to devising a 'scorched earth' or 'poison pill' strategy that strikes deeply at us but at some great cost to their own forces or population (a Pyrrhic victory).

2. *Proliferation.* Related to the previous issue, history also shows that innovations in military technologies—from armor and crossbows to intercontinental missiles and 'smart' bombs—give the inventing side a temporary advantage that is eroded over time by other nations working to replicate the technologies. Granting that modern technologies are more difficult to reverse-engineer or replicate than previous ones, it nevertheless seems inevitable or at least possible that they can be duplicated, especially if an intact sample can be captured, such as immobilizing a ground robot as opposed to shooting down an unmanned aerial vehicle. So with the development of autonomous military robots, we can expect their proliferation with other nations at some future point. This means that these robots—which we are currently touting as lethal, difficult-to-neutralize machines—may be turned against our own forces eventually.

   The proliferation of weapons, unfortunately, is an extremely difficult cycle to break: many nations are working to develop autonomous robotics, so a unilateral ban on their development would not accomplish much except to handicap that nation relative to the rest of the world. So the rush to develop this and other emerging technologies is understandable and irresistible, at least in today's world. One possible defense for our pursuit, apart from self-interested reasons, is that we (the US) want to ensure we develop these commanding technologies first, after which we would have more leverage to stop the proliferation of the same; further, because we occupy the higher moral ground, it would be most responsible for the US to develop the technologies first.

   The problem, of course, is that every nation thinks of itself as moral or 'doing the right thing', so it would be difficult to objectively assign a moral imperative to any given nation, including the US. Solving this problem, then, would seem to require additional legal and ethical theorizing, likely resulting in new international treaties and amendments to the Laws of War.

3. *Space race.* As on earth, autonomous robots may hold many benefits for space exploration [Jónsson et al., 2007]. Proliferation also has significant financial and environmental costs, particularly if military robotics technology is developed for outer space. First, launch costs are still astronomical, costing thousands of dollars *per pound* to put an object into low Earth orbit, and several times more per pound for geostationary orbit (not to mention periodic replacement costs and in-orbit repairs). An unlikely 'star wars' scenario aside—which would create countless pieces of space debris that would need to be tracked and threaten communications satellites and so on—even using robots for research purposes, e.g., to explore and develop moons and other planets, may spark another space race given the military advantages of securing the ultimate high ground. This not only opens up outer space for militarization, which the world's nations have largely resisted, but diverts limited resources that could make more valuable contributions elsewhere.

4. *Technology dependency*. The possibility that we might become dependent or addicted to our technologies has been raised throughout the history of technology and even with respect to robotics. Today, ethicists worry that we may become so reliant on, for instance, robots for difficult surgery that humans will start losing that life-saving skill and knowledge; or that we become so reliant on robots for basic, arduous labor that our economy is somehow impacted and we forget some of those techniques [Veruggio, 2007]. In the military, some soldiers already report being attached to the robot that saved their lives [Garreau, 2007].

   As a general objection to technology, this concern does not seem to have much force, since the benefits of the technology in question often outweigh any losses. For instance, our ability to perform mathematical calculations may have suffered somewhat given the inventions of the calculator and spreadsheets, but we would rather keep those tools even at that expense. Certainly, it is a possible hypothetical or future scenario that, after relying on robots to perform all our critical surgeries, some event—say, a terrorist attack or massive electromagnetic pulse—could interrupt an area's power supply, disabling our machines and leaving no one to perform the surgery (because we forgot how and have not trained surgeons on those procedures, since robots were able to do it better). But as abilities enhanced by technology, such as performing numeric calculations, have not entirely disappeared from a population or even to a life-impacting degree in individuals, it is unclear why we would expect something as artful as brain or heart surgery to be largely lost. Similarly, in the case of relying on robots for manual labor, technology dependency would not erase our ability to, say, dig holes to plant trees to any impacting degree.

5. *Civil security and privacy*. Defense technologies often turn into public or consumer technologies, as we previously pointed out. So a natural step in the evolution of military robots would seem to be their incarnation as civil security robots; they might guard corporate buildings, control crowds, and even chase down criminals. Many of the same concerns discussed above—such as technical challenges and questions of responsibility—would also become larger societal concerns: if a robot unintentionally (meaning that no human intentionally programmed it to ever do so) kills a small child, whether by accident (run over) or mistake (identification error), it will likely have greater repercussions than a robot that unintentionally kills a non-combatant in some faraway conflict. Therefore, there is increased urgency to address these military issues that may spill over into the public domain.

   And while we take it that soldiers, as government property, have significantly decreased privacy expectations and rights, the same is not true of the public at large. If and when robots are used more in society, and the robots are likely to be networked, concerns about illegal monitoring and surveillance—privacy violations—may again surface, as they have with most other modern

technologies, from DNA testing to genome sequencing to communications-monitoring software to nanotechnology. This raises the question of what kind of consent we need from the public before deploying these technologies in society.

7.6 Other and Future Challenges

1. *Co-opting of ethics effort by military for justification*. A possible challenge that does not fit neatly into any of the above categories is the following political concern. Defense organizations may be aware (now) of the above concerns, but they may still not choose to address the issues to mitigate risk by absolving themselves of this responsibility: they may simply point to ethicists and robot scientists working on related issues as justification for proceeding ahead without any real plan to address at least some of these risks [Sharkey, 2007b].

   This is an interesting *meta*-issue for robot ethics, i.e., it is about the study and aims of robot ethics and not so much about an issue directly related to the use of autonomous robots. While it is certainly a possibility that organizations may only pay 'lip-service' to the project of robot ethics to appease critics and watchdogs, it does not take much enlightenment or foresight to see actual, real-world benefits from earnestly addressing these challenges. Further, we might measure the commitment that organizations have to robot ethics by the funding levels for such research. And it would be readily apparent if, for instance, defense organizations ignored the counsel and recommendations of experts engaged in the field. This is to say that co-opting is a relatively transparent activity to identify, although the point is more that it could be too late (for those harmed or society in general) by then.

2. *Robot rights*. For now, robots are seen as merely a tool that humans use, morally no different (except in financial value) than a hammer or a rifle---their only value is instrumental, as a means to our ends. But as robots begin to assume aspects of human decision-making capabilities, the question may arise of their *intrinsic value*: do they deserve moral consideration of their own (beyond their financial or tactical value), and at what point in their evolution will they achieve this intrinsic value (as human lives seem to have)? When they become Kantian autonomous agents, making their own goals for themselves? Or would intrinsic value also require consciousness and emotions?

   Some technologists have suggested that, by 2029, robots will demand equal treatment before the law with humans—and believe that this demand will be granted [e.g., Kurzweil, 1999]. The only guarantee of avoiding this outcome appears to be a prohibition on programming robots with anything other than a 'slave morality', i.e., simply not allowing a Kantian-autonomous robot to ever be programmed or built (though such bans, especially when applied internationally, have

been notoriously difficult to enforce). It will require careful consideration in the future as to whether such a prohibition should ever be lifted. Fortunately, even 'technological optimists', such as Kurzweil, do not expect this to be an issue until at least the 2020s.

Thus far, we have not discussed the possibility of giving rights to robots, not so much that it is farfetched to do so (e.g., we give rights to non-living entities such as corporations) or to consider them as persons (philosophically-speaking; e.g., again corporations or ships or some animals such as dolphins), but that the prerequisites for rights seem to require advanced software or artificial intelligence that is not quite within our foreseeable grasp. Specifically, if our notion of personhood specifies that only persons can be afforded rights and that persons must have free will or the capacity for free will, then it is unclear whether we will ever develop technologies capable of giving free will or full autonomy to machines, and, indeed, we don't even know whether any other *biological* species will ever have or is now capable of such full autonomy; thus, we do not want to dwell on such a speculative issue here. That said, we will leave open the possibility that we may someday want or be logically required to give rights to robots [e.g., Kurzweil, 1999], but much more investigation is needed on the issue.

3. *The precautionary principle*. Given the above laundry list of concerns, some may advocate following a precautionary principle in robotics research—to slow or halt work until we have mitigated or addressed possible catastrophic risks—as critics have done for other technologies, such as bio- and nanotechnologies. For instance, those fearful of 'Terminator' scenarios where machines turn against us lesser humans, current research in autonomous robotics may represent a path towards possible, perhaps likely, disaster; thus a cautious, prudent approach would be to ban or at least significantly slow down research until we can sufficiently think about these issues before technology overtakes ethics. While we believe that a precautionary principle may be the appropriate course of action for some technology cases, many of the issues discussed above do not appear imminent enough to warrant a research moratorium or delay, just more investigation which may be sufficiently conducted in parallel to efforts to develop advanced robotics.

   Furthermore, a cautionary approach in the development of advanced systems is inherently in tension with both the approaches taken by the scientists and engineers developing robots and with the outlook of military planners, rapidly searching for more effective tools for the task of waging war. We will again leave open the possibility that someday we may have to seriously consider the role of the precautionary principle in robotics, but that day appears to be in the distant horizon and does not demand an extensive discussion here.

7.7 Further and Related Investigations Needed

Again, we do not intend the above to capture all possible issues related to autonomous military robotics. Certainly, new issues will emerge depending on how the technology and intended uses develop. In preceding sections, we have started to address what we seemed to be the most urgent and important issues to resolve first, especially as related to responsibility, risk, and the ability of robots to discriminate among targets. This is only the beginning of a dialogue in robot ethics and merits further investigations.

Moreover, our discussion here may be helpful in informing ethics research related to non-military robots, such as security, labor, and sex robots previously mentioned in the public domain. For instance, robots are already being used to care for the elderly, but are we merely pawning off our obligations to care for the elderly to machines who may be unable to provide the emotional content that seems to be needed in human relationships (even though there are significant advances in enabling robots to display 'emotions')? What are the benefits and risks of using robots as teachers, domestic help, or even as researchers, e.g., exploring difficult and alien environments? Do robotic planes, trains, and automobiles pose any special issues? These and other questions will need to be addressed; and as is often the case with public technologies, they have their roots in military innovations, so we may have a separate but related responsibility to begin looking ahead to these non-military questions as well.

8. Conclusions

> "We can only see a short distance ahead, but we can see plenty there that needs to be done."—Alan Turing [1950, p. 460]

There are many paths one may take in examining issues of risk and ethics arising from advanced military robotics. We initiate several lines of inquiry in this preliminary report, as follows.

In section 1, we begin by building the case for 'robot ethics.' While there are substantial benefits to be gained from the use of military robots, there are also many opportunities for these machines to act inappropriately, especially as they are given greater degrees of autonomy (for quicker, more efficient, and more accurate decision-making, and if they are to truly replace human soldiers). The need for robot ethics becomes more urgent when we consider pressures driving the market for military robotics as well as long-standing public skepticism that lives in popular culture.

In section 2, we lay the foundation for a robot-ethics investigation by presenting a wide range of military robots—ground, aerial, and marine—currently in use and predicted for the future. While most of these robots today are semi-autonomous (e.g., the US Air Force's Predator), some apparently-fully autonomous systems are emerging (e.g., the US Navy's Phalanx CIWS) though used in a very limited, last-resort context. From this, we can already see ethical questions emerge, especially related to the ability to discriminate combatants from non-combatants and the circumstances under which robots can make attack decisions on their own. We return to these questions in subsequent sections, particularly section 7.

In section 3, we look at behavioral frameworks that might ensure ethical actions in robots. It is natural to consider various programming approaches, since robots are related to our personal and business computer systems today that also depend on programmed instructions. We also recognize the forward-looking nature of our discussions here, given that the more-sophisticated programming abilities needed to build truly autonomous robotics are still under development.

We first discuss the traditional approach of *top-down programming*, i.e., establishing general rules that the robot would follow. A clear example is a deontological approach, such as using Kant's Categorical Imperative or Asimov's Laws of Robotics. However, a rigid set of rules is likely not robust enough to arrive at the correct action or decision in enough cases, particularly in unforeseen and complex scenarios. This suggests that we also need to attend to the 'rightness' of the result itself,

not just to the rules. But even if we acknowledge that consequences matter, there are other challenges raised by adopting a consequentialist/utilitarian approach, such as the impracticality of calculating and weighing all possible results, both near and far term, and the (strong) possibility of countenancing some intuitively-wrong action.

Given the apparent limitations of top-down programming, we then examine *bottom-up approaches*, inspired by biological evolution and human development. However, a key challenge is that bottom-up systems work best when they are directed at achieving one clear goal, but military robots often operate in dynamic environments in which available information is confusing or incomplete. That is, even if moral calculation is not an issue, there still remains the large problem of moral psychology, i.e., how to develop robots that embody the right tendencies in their reactions to the world and other agents in that world, particularly when the robots are confronted with a novel situation in which they cannot rely on experience.

Moral reasoning by humans, however, is not limited to exclusively a top-down or bottom-up approach; rather, we often use both strategies of rule-following and experience. (Nonetheless, it is useful to evaluate both programming approaches separately to identify their benefits and challenges.) Therefore, we consider a *hybrid approac*h of virtue ethic*s* for constructing ethical autonomous robots. This approach is concerned with the development of moral character; in the military case, with promoting the ideal character traits of a warfighter, i.e., a 'warrior code of ethics' as its virtues.

In section 4, we look at considerations in programming the Laws of War (LOW) and Rules of Engagement (ROE), which may differ from mission to mission, into a robot. No matter which programming approach is adopted, we at least would want to the robot to obey the LOW and ROE, and this might serve as a proxy for full-fledged morality until we have the capability to program a robot with the latter. Such an approach has several advantages, including: (1) any problems from moral relativism/particularism or other problems with general ethical principles are avoided; and (2) the relationship of morality to legality—a minefield for ethics—is likewise largely avoided, since the LOW and ROE make clear what actions are legal and illegal for robots, which serves as a reasonable approximation to the moral-immoral distinction.

Our discussion of the LOW and ROE, then, delves into their underlying foundation in just-war theory, particularly *jus ad bellum* (moral justification for entering war) and *jus in bello* (just and unjust actions in the prosecution of a war). We also examine ethical challenges to just-war theory as related to military robotics: Some have objected to the use of military robotics on the grounds that it makes easier the decision to enter war, in apparent violation of *jus ad bellum*; and we again see that the technical ability to properly discriminate against targets, as required by *jus in bello*, is a concern.

In section 5, we attend to the recurring possibility of accidental or unauthorized harm caused by robots; who would be responsible ultimately for those mishaps? We look at the issue through the lens of legal liability, both when robots are considered as merely products and when, as they are given more autonomy, they might be treated as legal agents, e.g., as legal quasi-persons such as children are regarded by the law. In the latter case, it is not clear how we would punish robots for their inappropriate actions.

In section 6, still attending to the possibility of unintended or unforeseen harm committed by a robot, we broaden our discussion by looking at how we might think about general risks posed by the machines and their acceptability. We offer a preliminary framework for a technology risk assessment, which includes the key factors of consent, informed consent, affected population, seriousness, and probability. This assessment highlights further the need for a lengthy period of rigorous testing and gradual rollout (crawl-walk-run approach) as a moral minimum for the responsible deployment of autonomous robots, especially by the military.

Finally, in section 7, we bring together a full range of issues raised throughout our examination, as well as some new issues, that must be recognized in any comprehensive assessment of risks from military robotics. These challenges fall into categories related to law, just-war theory, technical capabilities, human-robot interactions, general society, and other and future issues. For instance, we discuss such issues as:

- If a military robot refuses an order, e.g., if it has better situational awareness, then who would be responsible for its subsequent actions?
- How stringent should we take the generally-accepted 'eyes on target' requirement, i.e., under what circumstances might we allow robots to make attack decisions on their own?
- What precautions ought to be taken to prevent robots from running amok or turning against our own side, whether through malfunction, programming error, or capture and hacking?
- To the extent that military robots can help reduce instances of war crimes, what is the harm that may arise if the robots also unintentionally erode squad cohesion given their role as an 'outside' observer?
- Should robots be programmed to defend themselves—contrary to Arkin's position—given that they represent costly assets?
- Would using robots be counterproductive to winning the hearts and minds of occupied populations or result in more desperate terrorist-tactics given an increasing asymmetry in warfare?

From the preceding investigation, we can draw some general and preliminary conclusions, including some future work needed:

1. Creating autonomous military robots that can act *at least as* ethically as human soldiers appears to be a sensible goal, at least for the foreseeable future and in contrast to a greater demand of a perfectly-ethical robot. However, there are still daunting challenges in meeting even this relatively-low standard, such as the key difficulty of programming a robot to reliably distinguish enemy combatants from non-combatants, as required by the Laws of War and most Rules of Engagement.

2. While a faster introduction of robots in military affairs may save more lives of human soldiers and reduce war crimes committed, we must be careful to not unduly rush the process. Much different than rushing technology products to commercial markets, design and programming bugs in military robotics would likely have serious, fatal consequences. Therefore, a rigorous testing phase of robots is critical, as well as a thorough study of related policy issues, e.g., how the US Federal Aviation Administration (FAA) handles UAVs flying in our domestic National Airspace System (which we have not addressed here).

3. Understandably, much ongoing work in military robotics is likely shrouded in secrecy; but a balance between national security and public disclosure needs to be maintained in order to help accurately anticipate and address issues of risk or other societal concerns. For instance, there is little information on US military plans to deploy robots in space, yet this seems to be a highly strategic area in which robots can lend tremendous value; however, there are important environmental and political sensitivities that would surround such a program.

4. Serious conceptual challenges exist with the two primary programming approaches today: top-down (e.g., rule-following) and bottom-up (e.g., machine learning). Thus a hybrid approach should be considered in creating a behavioral framework. To this end, we need to a clear understanding of what a 'warrior code of ethics' might entail, if we take a virtue-ethics approach in programming.

5. In the meantime, as we wait for technology to sufficiently advance in order to create a workable behavioral framework, it may be an acceptable proxy to program robots to comply with the Laws of War and appropriate Rules of Engagement. However, this too is much easier said than done, and at least the technical challenge of proper discrimination would persist and require resolution.

6. Given technical limitations, such as programming a robot with the ability to sufficiently discriminate against valid and invalid targets, we expect that accidents will continue to occur, which raise the question of legal responsibility. More work needs to be done to clarify the chain of responsibility in both military and civilian contexts. Product liability laws

are informative but untested as they relate to robotics with any significant degree of autonomy.

7. Assessing technological risks, whether through the basic framework we offer in section 6 or some other framework, depends on identifying potential issues in risk and ethics. These issues vary from: foundational questions of whether autonomous robotics can be legally and morally deployed in the first place, to theoretical questions about adopting precautionary approaches, to forward-looking questions about giving rights to truly autonomous robots. These discussions need to be more fully developed and expanded.

8. Specifically, the challenge of creating a robot that can properly discriminate among targets is one of the most urgent, particularly if one believes that the (increased) deployment of war robots is inevitable. While this is a technical challenge and resolvable depending on advances in programming and AI, there are some workaround policy solutions that can be anticipated and further explored, such as: limiting deployment of lethal robots to only inside a 'kill box'; or designing a robot to target only other machines or weapons; or not giving robots a self-defense mechanism so that they may act more conservatively to prevent; or even creating robots with only non-lethal or less-than-lethal strike capabilities, at least initially until they are proven to be reliable.

These and other considerations warrant further, more detailed investigations in military robotics and issues of design, risk, and ethics. Such interdisciplinary investigations will require collaboration among policymakers and analysts, roboticists, ethicists, sociologists, psychologists, and others, internationally and including the general public as a key stakeholder. And this work has the potential to be as broad as other fields in science and society, such as bioethics or computer ethics.

The use of military robots represents a new era in warfare, perhaps more so than crossbows, airplanes, nuclear weapons, and other innovations have previously. Robots are not merely another asset in the military toolbox, but they are meant to also replace human soldiers, especially in 'dull, dirty, and dangerous' jobs. As such, they raise novel ethical and social questions that we should confront as far in advance as possible—particularly before irrational public fears or accidents arising from military robotics derail research progress and national security interests.

9. References

Allen, Colin, Varner, Gary, and Zinser, Jason (2000). "Prolegomena to Any Future Artificial Moral Agent", *Journal of Experimental and Theoretical Artificial Intelligence* 12.3:251–261.

Anderson, Michael, and Anderson, Susan Leigh (2007). "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* 28.4: 15–26.

Arkin, Ronald C. (1998). *Behavior-Based Robotics,* Cambridge: MIT Press.

Arkin, Ronald C. (2007). *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Hybrid Robot Architecture*, Report GIT-GVU-07-11, Atlanta, GA: Georgia Institute of Technology's GVU Center. Last accessed on September 15, 2008: http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf

Asaro, Peter (2007). "Robots and Responsibility from a Legal Perspective", Proceedings of the IEEE 2007 International Conference on Robotics and Automation, Workshop on RoboEthics, April 14, 2007, Rome, Italy. Last accessed on September 15, 2008: http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf

Asaro, Peter (2008). "How Just Could a Robot War Be?" in Adam Briggle, Katinka Waelbers, and Philip Brey (eds.) *Current Issues in Computing and Philosophy*, pp. 50-64, Amsterdam, The Netherlands: IOS Press.

Asimov, Isaac (1950). *I, Robot* (2004 edition), New York, NY: Bantam Dell.

Asimov, Isaac (1957). *The Naked Sun*, New York, NY: Doubleday.

Asimov, Isaac (1985). *Robots and Empire*, New York, NY: Doubleday.

BBC (2005). "SLA Confirm Spy Plane Crash", *BBC.com*, October 19, 2005. Last accessed on September 15, 2008: http://www.bbc.co.uk/sinhala/news/story/2005/10/051019_uav_vavunia.shtml

BBC (2007). "Robotic Age Poses Ethical Dilemma", *BBC.com*, March 7, 2007. Last accessed on September 15, 2008: http://news.bbc.co.uk/2/hi/technology/6425927.stm

Bekey, George (2005). *Autonomous Robots: From Biological Inspiration to Implementation and Control*, Cambridge, MA: MIT Press.

Brooks, Rodney (2002). *Flesh and Machines*. New York: Pantheon Books.

Canning, John, Riggs, G.W., Holland, O. Thomas, Blakelock, Carolyn (2004). "A Concept for the Operation of Armed Autonomous Systems on the Battlefield", *Proceedings of Association for Unmanned Vehicle Systems International's (AUVSI) Unmanned Systems North America*, August 3-5, 2004, Anaheim, CA.

Canning, John (2008). "Weaponized Unmanned Systems: A Transformational Warfighting Opportunity, Government Roles in Making it Happen", 2008 American Society of Naval Engineers' (ASNE) Proceedings of Engineering the Total Ship (ETS) Symposium, September 23-25. 2008, Falls Church, VA.

Čapek, Karel (1921). *R.U.R.* (2004 edition, trans. Claudia Novack), New York, NY: Penguin Group.

CBS (2007). "Robots Playing Larger Role in Iraq War", October 21, 2007 news report. Last accessed on September 15, 2008: http://cbs3.com/topstories/robots.iraq.army.2.410518.html

Churchland, Paul (1995). *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*, Cambridge, MA: MIT Press.

Clarke, Roger (1994). "Asimov's Laws of Robotics: Implications for Information Technology," *IEEE Computer* (part 1: December 1993, pp. 53–61; part 2: January 1994, pp. 57–66).

Coffee, Jr., John C. (1981). "'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment," *Michigan Law Review*, Vol. 79, No. 3, pp. 386-459.

Computer Professionals for Social Responsibility (2008). "Technology in Wartime" conference, January 26, 2008, Stanford, CA. Last accessed on September 15, 2008: http://technologyinwartime.org/

Davis, Burke (1980). *Sherman's March: The First Full-Length Narrative of General William T. Sherman's Devastating March through Georgia and the Carolinas*, New York, NY: Random House.

DeMoss, David (1998). "Aristotle, Connectionism, and the Morally Excellent Brain", *The Proceedings of the Twentieth World Congress of Philosophy*, August 10-15, 1998, Boston, MA.

DesJardins, Joseph (2003). *An Introduction to Business Ethics*, pp. 99-103, Columbus, OH: McGraw-Hill.

Dilov, Lyuben (1974). *The Way of Icarus*. Дилов, Любен. Пътят на Икар. Захари Стоянов. ISBN 954-739-338-3.

Fikes, Richard and Nilsson, Nils (1971). "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving", *Artificial Intelligence* 2(3-4): 189-208.

Foot, Phillipa (1972). "Morality as a System of Hypothetical Imperatives", *The Philosophical Review* 81.3: 305-316.

Garreau, Joel (2007). "Bots on the Ground", *Washington Post*, May 6, 2007. Last accessed on September 15, 2008: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009_pf.html

Harrison, Harry (1989). :The Fourth Law of Robotics", in Isaac Asimov and Martin Harry Greenberg (eds.) *Foundation's Friends: Stories in Honor of Isaac Asimov*, New York, NY: Tor Books.

Hemingway, Ernest (1935). "Notes on the Next War: A Serious Topical Letter", *Esquire*, Vol. 4, No. 3: 19, 156.

Hew, Patrick (2007). "Autonomous Situation Awareness: Implications for Future Warfighting", *Australian Defence Force Journal* 174: 71-87. Last accessed on September 15, 2008: http://www.defence.gov.au/publications/dfj/index.htm

Hobbes, Thomas (1651). *Leviathan* (1982 edition), New York, NY: Penguin Group.

Institute of Electrical and Electronics Engineers (2008). "International Conference on Advanced Robotics and its Social Impact" conference, August 23-25, 2008, Taipei, Taiwan. Last accessed on September 15, 2008: http://arso2008.ntu.edu.tw/

International Federation of Robotics (2008). "International Robot Standards" page from International Federation of Robotics website. Last accessed on September 15, 2008: http://www.ifr.org/modules.php?name=News&file=article&sid=20

Iraq Coalition Casualty Count (2008). "Deaths Caused by IEDs" and "U.S. Deaths by Month" webpages. Last accessed on September 15, 2008: http://icasualties.org/oif/IED.aspx and http://icasualties.org/oif/USDeathByMonth.aspx

Johnson, Robert (2008). "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy, Fall 2008 Edition. Last accessed on September 15, 2008: http://plato.stanford.edu/archives/fall2008/entries/kant-moral/

Jónsson, Ari, Morris, Robert, and Pedersen, Liam (2007). "Autonomy in Space: Current Capabilities and Future Challenges", *AI Magazine* 28.4: 27-42.

Joy, Bill (2000). "Why the Future Doesn't Need Us", *Wired* 8.04: 238-262.

Kahn, Paul (2002). "The Paradox of Riskless War," *Philosophy & Public Policy Quarterly*, Vol. 22: 2–8.

Kant, Immanuel (1785). *Grounding for the Metaphysics of Morals* (1993 edition, translated by James W. Ellington), Indianapolis, IN: Hackett Publishing Co.

Kurzweil, Ray (1999). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York, NY: Viking Penguin.

Kurzweil, Ray (2005). *The Singularity is Near: When Humans Transcend Biology*, New York, NY: Viking Penguin.

Lee, Steven (2004). "Double Effect, Double Intention, and Asymmetric Warfare", *Journal of Military Ethics*, 3.3: 233-251.

Levy, David (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*, New York, NY: HarperCollins Publishers.

McIntyre, Alison (2004). "Doctrine of Double Effect", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition). Last accessed on September 15, 2008: http://plato.stanford.edu/archives/fall2008/entries/double-effect/

Murray, Mary Elizabeth (2008). "Moral Development and Moral Education: An Overview", University of Illinois at Chicago website. Last accessed on September 15, 2008: http://tigger.uic.edu/~lnucci/MoralEd/overview.html

National Defense Authorization Act (2000). Floyd D. Spence National Defense Authorization Act for Fiscal Year 2001, Public Law 106-398, Section 220. Last accessed on September 15, 2008: http://www.dod.mil/dodgc/olc/docs/2001NDAA.pdf

National Transportation Safety Board (2007). "NTSB Cites Wide Range of Safety Issues in First Investigation of Unmanned Aircraft Accident", NTSB press release, October 16, 2007. Last accessed on September 15, 2008: http://www.ntsb.gov/Pressrel/2007/071016b.htm

North American Computing and Philosophy (2008). "The Limits of Computation" conference, July 10-12, 2008, Bloomington, IN. Last accessed on September 15, 2008: http://www.ia-cap.org/na-cap08/index.htm

O'Brien, William V. (1981). *The Conduct of Just and Limited War,* New York, NY: Praeger Publishers.

Oh, Daniel (2008). "The Relevance of Virtue Ethics and Application to the Formation of Character Development in Warriors", *The Army Chaplaincy* online journal, Spring-Summer 2008. Last accessed on September 15, 2008: http://www.usachcs.army.mil/TACarchive/tacss08/tacss08oh7.pdf

Orend, Brian (2001). *Michael Walzer on War and Justice*, Montreal, Quebec: McGill-Queen's University Press.

Orend, Brian (2002). "Justice After War", *Ethics & International Affairs* 16.1: 43-56.

Orend, Brian (2006). *The Morality of War*, Peterborough, Ontario: Broadview Press.

Padgett, Tim (2008). "Florida's Blackout: A Warning Sign?", *Time.com*, February 27, 2008. Last accessed on September 15, 2008: http://www.time.com/time/nation/article/0,8599,1717878,00.html

Page, Lewis (2008). "US War Robots 'Turned Guns' on Fleshy Comrades", *The Register* (UK), April 11, 2008. Last accessed on September 15, 2008: http://www.theregister.co.uk/2008/04/11/us_war_robot_rebellion_iraq/

Royal United Services Institute (RUSI) for Defence and Security Studies (2008). "The Ethics of Autonomous Military Systems" conference, February 27, 2008, London, UK. Last accessed on September 15, 2008: http://www.rusi.org/events/past/ref:E47385996DA7D3/

Rowe, Neil C. (2008). "Ethics of Cyber War Attacks", in Lech J. Janczewski and Andrew M. Colarik (eds.) *Cyber Warfare and Cyber Terrorism*, Hershey, PA: Information Science Reference

Russell, Stuart J., and Norvig, Peter (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, NJ: Prentice Hall.

Searle, John (1980). "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3.3: 417-457

Shachtman, Noah (2007). "Robot Cannon Kills 9, Wounds 14", *Wired.com,* October 18, 2007. Last accessed on September 15, 2008: http://blog.wired.com/defense/2007/10/robot-cannon-ki.html

Sharkey, Noel (2007a). "Robot Wars are a Reality", *The Guardian* (UK), August 18, 2007, p. 29. Last accessed on September 15, 2008: http://www.guardian.co.uk/commentisfree/2007/aug/18/comment.military

Sharkey, Noel (2007b). "Automated Killers and the Computing Profession", *Computer* 40: 122-124. Last accessed on September 15, 2008: http://www.computer.org/portal/site/computer/menuitem.5d61c1d591162e4b0ef1bd108bcd45f3/index.jsp?&pName=computer_level1_article&TheCat=1015&path=computer/homepage/Nov07&file=profession.xml&xsl=article.xsl&

Sharkey, Noel (2008a). "Cassandra or False Prophet of Doom: AI Robots and War", *IEEE Intelligent Systems*, July/August 2008, pp. 14-17. Last accessed on September 15, 2008: http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2008/X4-08/x4his.pdf

Sharkey, Noel (2008b). "Grounds for Discrimination: Autonomous Robot Weapons", *RUSI Defence Systems*, 11.2: 86-89.

Simon, Herbert (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (1997 fourth edition), New York, NY: Free Press.

Sofge, Erik (2008). **"**The Inside Story of the SWORDS Armed Robot 'Pullout' in Iraq: Update", *PopularMechanics.com*, April 15, 2008. Last accessed on September 15, 2008: http://www.popularmechanics.com/blogs/technology_news/4258963.html

Solomon, David (1988). "Internal Objections to Virtue Ethics", in Peter A. French, Theodore Uehling, Jr., and Howard Wettstein (eds.), *Midwest Studies in Philosophy Vol. XIII Ethical Theory: Character and Virtue*, Notre Dame, IN: University of Notre Dame Press.

Solum, Lawrence (1992). "Legal Personhood for Artificial Intelligences," *North Carolina Law Review*, Vol. 70: 1231-1287.

Sparrow, Rob (2007). "Killer Robots", *Journal of Applied Philosophy*, Vol. 24, No. 1: 62-77

Thompson, Paul B. (2007). *Food Biotechnology in Ethical Perspective, 2nd ed.*, Dordrecht, The Netherlands: Springer.

Turing, Alan (1950). "Computing Machinery and Intelligence", *Mind,* Vol. 59, No. 236: 434–460.

University of San Diego (2008). Ethics Updates webpage. Last accessed on September 15, 2008: http://ethics.sandiego.edu/index.asp

US Army Surgeon General's Office (2006). *Mental Health Advisory Team (MHAT) IV: Operation Iraqi Freedom 05-07,* November 16, 2006. Last accessed on September 15, 2008: http://www.globalpolicy.org/security/issues/iraq/attack/consequences/2006/1117mhatreport.pdf

US Army Surgeon General's Office (2008). *Mental Health Advisory Team (MHAT) V: Operation Iraqi Freedom 06-08,* February 14, 2008. Last accessed on September 15, 2008: http://www.armymedicine.army.mil/reports/mhat/mhat_v/Redacted1-MHATV-OIF-4-FEB-2008Report.pdf

US Department of Defense (2007). *Unmanned Systems Roadmap 2007-2032*, Washington, DC: Government Printing Office. Last accessed on September 15, 2008: http://www.acq.osd.mil/usd/Unmanned%20Systems%20Roadmap.2007-2032.pdf

US Department of Energy (2004). *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, Washington, DC: Government Printing Office. Last accessed on September 15, 2008: https://reports.energy.gov/BlackoutFinal-Web.pdf

US Department of the Navy (2004). *Navy Unmanned Undersea Vehicle (UUV) Master Plan*, Washington, DC: Government Printing Office. Last accessed on September 15, 2008: http://www.navy.mil/navydata/technology/uuvmp.pdf.

Van der Loos, H.F. Machiel (2007). "Ethics by Design: A Conceptual Approach to Personal and Service Robot Systems", *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethic*s, April 14, 2007, Rome, Italy.

Veruggio, Gianmarco (2007). *EURON Roboethics Roadmap*, Genova, Italy: European Robotics Research Network. Last accessed on September 15, 2008:

http://www.roboethics.org/icra07/contributions/VERUGGIO%20Roboethics%20Roadmap%20Rel.1.2.pdf

Wallach, Wendell and Allen, Colin (2008). *Moral Machines: Teaching Robots Right from Wrong*, New York, NY: Oxford University Press.

Walter, W.G. (1950). "An Imitation of Life", *Scientific American,* 182:42-45.

Walzer, Michael (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, New York, NY: Basic Books.

Weckert, John, ed. (2007). *Computer Ethics*, Burlington, VT: Ashgate Publishing.

Appendix A: Definitions

While we do not want to be entangled with debating precise definitions in this report, it nevertheless would be useful to give more detailed explanations of our key terms—namely 'robot', 'autonomy', and 'ethics'—to ensure a common understanding from the start:

A.1 Robot

Before we offer a working definition of a robot, let us note some historical origins: In 1921, the world was introduced to modern concept of a robot in the popular play *R.U.R.* (or *Rossum's Universal Robots*) by Czech author Karel Čapek. The dystopian play featured factory-built, artificial people, who can be mistaken as humans, called robots—whose namesake is derived from the Czech word 'robota' which means 'drudgery' or 'servitude' or 'labor', and these engineered slaves ultimately rebel against their human masters.

Čapek's robots were biological based and more akin to the resurrected man-creature in Mary Shelley's *Frankenstein* or the replicants of Ridley Scott's *Blade Runner* than to the modern fusion of computer and machine, popularized by science-fiction author Isaac Asimov and others. Therefore, the idea that robots are electromechanical is not part of the original conception of robots; nor does it seem to be an essential feature, since we can imagine advances in genetic engineering and synthetic biology to some day enable us to create biological-based artificial creatures that function as today's robots do and can be called 'robots' (or called something else if such developments cause us to evolve our terminology, as 'robot' had replaced older words such as 'automaton' and 'android'). That said, we will concern ourselves here primarily with electromechanical machines, though we will leave open the possibility of biological and virtual machines or creations as robots.

Further, though the original notion of a robot is tied with automation of work, we are more interested in how the definition of robot differentiates a robot from a garden-variety, mere machine. That is, our definition is not meant to help elucidate the difference between a robot and automation or work, rather to help explain why objects such as a laptop computer or a coffee machine do not count as robots.

To the definition now, in its most basic sense, we define '**robot**' as a *machine that senses, thinks, and acts*: "Thus a robot must have sensors, processing ability that emulates some aspects of cognition,

and actuators. Sensors are needed to obtain information from the environment. Reactive behaviors (like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment. Generally, these forces will result in motion of the entire robot or one of its elements (such as an arm, a leg, or a wheel)" [Bekey, 2005].

For all practical purposes today, this means that a robot is essentially a computer with sensory inputs (that do not require direct and intentional human action, such as required by a keyboard or touchpad; i.e., it can direct itself given certain environmental inputs) and non-digital output (i.e., more than manipulation of data or pixels or even sending a file to printer, but an ability to move some part of itself in order to manipulate real-world objects). This is neither a precise definition nor description, and it certainly needs to be refined; but it at least begins to meet our needs for a general understanding of what counts and what does not count as a robot.

A standard, more exact definition, however, proves elusive, as perhaps evidenced by the fact that no major robotics organization to our knowledge—including Robotic Industries Association, IEEE and its Robotics & Automation Society, European Robotics Research Network, Japan Robot Association, Australian Robotics & Automation Association, and others—provides a clear definition of the term on their respective sites or publications, as far as we can tell. Some organizations, such as International Federation of Robotics, follow the definition given by the International Organization for Standardization (ISO) for *manipulating industrial robots*: "an automatically controlled, reprogrammable, multipurpose, manipulator programmable in three or more axes, which may be either fixed in place or mobile for use in industrial automation applications" [International Federation of Robotics, 2008]. But this is far from an adequate definition for a generic robot, and no ISO definition for such a robot has been devised, to our knowledge.

In a military context, we have a more useful understanding from the US Department of Defense's (DoD) definition of an 'unmanned vehicle' (though some of these might not properly be robots, at least under our working definition): "A powered vehicle that does not carry a human operator, can be operated autonomously or remotely, can be expendable or recoverable, and can carry a lethal or non-lethal payload. Ballistic or semi-ballistic vehicles, cruise missiles, artillery projectiles, torpedoes, mines, satellites, and unattended sensors (with no form of propulsion) are not considered unmanned vehicles. Unmanned vehicles are the primary component of unmanned systems" [US Department of Defense, 2007, p. 1].

We take 'vehicle' to mean a mobile, maneuverable machine such as a car, boat, or airplane, but in our analysis, a robot need not be mobile (though it seems most will be in military applications). Yet some degree of mobility is an essential feature of a robot, as we mentioned in our first definition of a robot above; for instance, a robot can be a fixed manufacturing machine with movable arms or an

immobile sentry robot with swiveling gun turrets. The mobility requirement here has less to do with moving from one physical location to another (although most robots can and will do this) but more with the ability to interact with and manipulate the external, physical world to some meaningful degree. This requirement then differentiates a robot from, say, a computer with environmental sensors but that can only run software programs and not exert a sufficient amount of force on the outside world (beyond spitting out pages of paper or opening its CD player door).

Relatedly, the concept that the machine is *powered* is needed in a sensible definition of a robot. Though it is taken for granted that all machines operate under some power, especially where mobility and information processing is required, we want to rule out 'dumb' machines that are not internally driven but nonetheless seem to sense, think, and act. For instance, a small sensor may be designed to be carried by wind or water and mechanically perform some action under certain conditions (such as release a payload when a thermal, chemical, or gyroscopic switch is tripped): such a machine appears to sense, act, and think (to the extent it performs an action under the right conditions, similar to a mechanical calculator of the early twentieth century), yet we would resist calling it a robot, though it would still be considered a machine, i.e., robots belong to a subset of machines. This is not to say that very small robots cannot be carried by wind or waves, but they would also need to convert or otherwise use that energy, or use some other power, to independently move itself or some part of the machine. Thus, our definition of a robot should include the notion of *internal or self-directed power* (e.g., electricity generated by a battery or harnessed from solar or wave energy), as well as the existence of something to be powered (e.g., a computer processing chip or payload-release mechanism).

However, the DoD definition needs to be modified for our use in this report: It seems to be overly broad to include fully-remote-controlled machines as robots, since many children's toys would qualify as robots, such as a toy car tethered or wirelessly connected to a control knob (though a few toys, such as AIBO™ or Pleo™ or Robosapien™, may truly be robots). That is, most of these toys do not make decisions for themselves; they depend on a human actor. Rather, the generally-accepted idea of a robot depends critically on the notion that it has some degree of autonomy or can 'think' for itself, as it makes decisions and acts upon the environment. Thus, the Air Force's Predator, though mostly tele-operated by humans, makes some navigational decisions on its own and therefore would count as a robot. Further, robots need not be unmanned, though many are and will be. It is conceivable that a robot may indeed be partially autonomous *and* carry a human operator who makes some decisions, so we do not want to rule such a machine out in our definition. Indeed, beyond the distinction between a robot and a mere machine, the line between robot and human may soon become blurred as robotic technologies are integrated with biological bodies.

As for the ISO's requirements of multi-purpose, reprogrammable, and movable on three or more axes, those seem to be unnecessarily limiting for a general definition of a robot, though perhaps

appropriate for the ISO's purpose of defining a manipulating industrial robot. So we will not include those requirements in our conception of a robot here, though further investigations may warrant them. For instance, we are leaving open the question whether a robot needs to be programmable or reprogrammable, since we may envision counterexamples of a disposable, one-time-use robotic insect that is not reprogrammable or some autonomous robot that transcends its programming.

Therefore, our working definition of a robot—a powered machine that (1) senses, (2) thinks (in a deliberative, non-mechanical sense), and (3) acts—appears defensible and comprehensive. In addition to the cases preciously discussed, it can rule out as robots the following (current but perhaps not future) types of ordnances and technologies: ballistic or semi-ballistic vehicles, cruise missiles, artillery projectiles, torpedoes, mines, satellites, and unattended sensors (with no form of propulsion). However, a 'smart' mine, for instance, conceivably may be developed in the future such that it can sense, think (i.e., discriminate among targets), and act (i.e., exert forces upon the environment, other than self-destruct) and therefore considered to be a robot. In non-military contexts, our definition eliminates mundane objects, such as coffee machines (they don't think; see related discussion below about autonomy) and personal computers (which, by themselves, don't exert an influence on or sense the external world in a significant way, and they require human inputs), as robots.

A.2 Autonomy

This brings us to another critical concept we need to define: autonomy. Though this task is even more difficult than the former, we will offer less analysis, given that the different conceptions, complexity, and applications of autonomy are well covered throughout philosophical and legal literature (and such discussions about the definition of a robot has not been nearly extensively covered in technical or other literature).[8] For the purposes of this report, it will suffice to initially stipulate '**autonomy**' to be about *the capacity to operate in the real-world environment without any form of external control for extended periods of time* [Bekey, 2005]. (But see the refined definition below.)

[8] We want to recognize a *technical* account of autonomy, such that autonomy is measured by the amount of time it takes for a system to refer or 'check back' with a human before proceeding with a certain action [Hew, 2007]. Thus, a landmine has infinite autonomy, since it never needs to refer back to a human for authorization once it has been armed; and a fully remote-controlled robot would have no autonomy, since it is constantly referring back to a human (and indeed completely dependent on a human) for instructions. While this may be a useful account of autonomy in some technical discussions, it does not seem to be relevant to a discussion about ethics and risk, to the extent that such issues arise from a system's ability to make unpredicted, unforeseen, unanticipated, or undesirable choices. Further, from a legal and ethical standpoint, it seems to be incoherent to ascribe autonomy to unthinking objects such as landmines or a toaster.

This is to say that, in defining the term, we are not interested at this point in issues traditionally linked to autonomy, such as the assignment of political rights and moral responsibility (as different from legal responsibility) or even more philosophical issues related to free will, moral agency, personhood, and whether machines can even 'think' and have intentions (as opposed to merely being programmed to achieve some goal)—as important as those issues are in philosophy, law, and ethics. Therefore, in the interest of simplicity, we will content ourselves to define and discuss autonomy in the context of human-created machines.

The notion of autonomy is important to help elucidate the second criteria of 'thinking' in our basic working definition of a robot. Like autonomy, much controversy surrounds our understanding of 'thinking', especially whether it is appropriate to apply that term to machines. By this term, we do not mean the mere capability of information or data processing; that would make most of our electronic devices into thinking things and make the term overly broad. Rather, by 'thinking', we mean to include some degree of autonomy or decision-making not influenced by external controllers, giving the machine an appearance of deliberative thought, if not the actual ability to meaningfully make choices.[9]

Thus, given our initial definition of autonomy, *fully* remote- or tele-operated machines would not count as autonomous, since they are not operated without external control; they cannot 'think' and therefore cannot act for themselves. (Again, tele-operated vehicles such as the Air Force's Predator would count as robots under our working definition, because they have some autonomy, such as in navigation, even if they do not make any strike decisions.) Neither are today's desktop or laptop computers autonomous, since they still require human inputs. Yet a problem with our simple definition may arise: wouldn't autonomous robots simply be sensing and moving computers that run programs, like everyday computers; and through these programs, both computers and robots can be said to be 'externally controlled' by the programmer or team of programmers that created the program? That is to say, the notion of *external control* is vague and begs for clarification.

Let's retreat one step to ask the following: might we consider some computers as *semi*-autonomous, such as one that runs a computer program that enables an avatar (or virtual-reality character or persona) to run without further external control, i.e., an avatar that seems to act on its own, as some already do now? Surely, such an avatar would be considered at least semi-autonomous, at least by popular standards, but how do we adjust our definition of autonomy to include such a case?

[9] A logical implication of this is that *all* robots, as we define them, will have some degree of autonomy, making 'autonomous robot' a redundant expression; but we will nevertheless keep with this expression to signal that we are referring to robots that have a greater degree of autonomy than usual, if one considers autonomy as a spectrum from unthinking automatons (such as bacteria and simple organisms) to semi-autonomous beings (such as children, some animals, and some robots today) to fully-autonomous, moral agents.

While it is true that programs need to be created by some programmer, even programs written by other programs, there will always be some external, human cause for whatever actions machine exhibit; so artificial autonomy would be impossible if no person can play a role in the causal chain, especially at the programming level. Indeed, the philosophical position that free will does *not* exist seems to depend on a related argument, that in a scientific, deterministic world, there must be some prior, external cause for our behavior that we are not responsible for (and thus we are not responsible for our consequent actions or even have the power to alter that chain of events).

So to avoid this problem, let us stipulate that artificial autonomy is possible and that it does *not* imply that a machine's actions are undetermined or unpredictable (as may be required in the usual conception of free will). This seems to require that we exempt programmers and designers from the considered causal chain, such that a robot running a sophisticated program may be considered to be semi- or fully-autonomous, even though its actions may be predetermined given certain environmental conditions, at the programming or design level. Thus, we may also consider some computers today to be semi-autonomous in letting loose a self-directed avatar upon the virtual world.

Given our use of 'autonomy' here, a semi- or fully-autonomous robot would be able to choose to perform at least some actions 'on its own' or without a human determining—at least not at a design or programming level—what course of action it should take. Thus to refine our initial definition, we take **autonomy** in machines to mean: *the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.*

A.3 Ethics

Finally, in this report, we use the term '**ethics**' broadly to include not just normative issues, i.e., questions about what we should or ought to do, but also general concerns related to social and cultural impact as well as risk, e.g., responsibility after a malfunction, arising from the use of robotics. As a result, we will cover all these areas in our report, not just philosophical questions or ethical theory, with the goal of providing some relevant if not actionable at this preliminary stage. (This diverges from the traditional, more narrow conception of ethics, at least as understood in academic philosophy.)

A note on the issue of ethics: Is robot ethics a subset of military ethics, or computer ethics, or some other area of ethics? We believe that robot ethics is emerging to become a field unto its own. There is still a critical gap between machine or computer ethics and military ethics, in which important questions are only now being raised about the moral responsibility, risk, and just use in war of

relatively autonomous systems (robots or computer networks). Further, robot ethics is not limited to military-related issues; there are new dilemmas related to the use of robots as a proxy for elderly care, for sexual relationships, for human workers especially those with specialized skills such as surgeons, and so on. Thus, the introduction of 'smart' robotics into the military and the marketplace has implications for both military and other ethics, with potentially transformative consequences on many traditional issues, such as the nature of personhood, agency, autonomy, and even what it means to be a soldier.

Appendix B: Contacts

1. **Patrick Lin, Ph.D.**
   California Polytechnic State University
   Ethics + Emerging Sciences Group
   Philosophy Department
   1 Grand Avenue
   Building 47, Room 37
   San Luis Obispo, California 93407
   Email: palin@calpoly.edu
   Dept. phone: 805-756-2041

2. **George Bekey, Ph.D.**
   California Polytechnic State University
   Ethics + Emerging Sciences Group
   Biomedical/General Engineering Department
   Building 13, Room 260
   San Luis Obispo, California 93407
   email: gbekey@calpoly.edu
   Dept. phone: 805-756-6400

3. **Keith Abney, M.A.**
   California Polytechnic State University
   Ethics + Emerging Sciences Group
   Philosophy Department
   1 Grand Avenue
   Building 47, Room 37
   San Luis Obispo, California 93407
   email: kabney@calpoly.edu
   Dept. phone: 805-756-2041

This work is sponsored by the Department of the Navy, Office of Naval Research, under awards # N00014-07-1-1152 and N00014-08-1-1209.