Full Title: The Case for Ethical Autonomy in Unmanned Systems

Running Title: Ethical Autonomous Systems

Author: Ronald C. Arkin

Contact: Email: arkin@cc.gatech.edu
Telephone: 1-404-894-9311

Institutional Affiliation and Address:

Georgia Institute of Technology
85 5th Street NW
Atlanta, Georgia, U.S.A. 30308

Biographical Sketch

Ronald C. Arkin is Regents' Professor of Computer Science and Director of the Mobile Robot Laboratory at Georgia Tech University (Atlanta, GA). He also serves as the Associate Dean for Research and Space Planning in the College of Computing at Georgia Tech since October 2008. During 1997-98, Professor Arkin served as STINT visiting Professor at the Centre for Autonomous Systems at the Royal Institute of Technology (KTH) in Stockholm, Sweden. From June-September 2005, Prof. Arkin held a Sabbatical Chair at the Sony Intelligence Dynamics Laboratory in Tokyo, Japan and then served as a member of the Robotics and Artificial Intelligence Group at LAAS/CNRS in Toulouse, France from October 2005-August 2006. His research interests include behavior-based reactive control and action-oriented perception for mobile robots and unmanned aerial vehicles, hybrid deliberative/reactive software architectures, robot survivability, multiagent robotic systems, biorobotics, human-robot interaction, robot ethics, and learning in autonomous systems. He has over 170 technical publications in these areas. Prof. Arkin has written a textbook entitled *Behavior Based Robotics*, published hed by MIT Press in May 1998, co-edited (with G. Bekey) a book entitled *Robot Colonies* published in 1997, and a book published in Spring 2009 entitled *Governing Lethal Behavior in Autonomous Robots* published by Chapman-Hall (Taylor & Francis). The following article is derived principally from that latter work.

Abstract

The underlying thesis of research in ethical autonomy for lethal autonomous unmanned systems is that they will potentially be capable of performing more ethically on the battlefield than are human soldiers. In this article this hypothesis is supported by ongoing and foreseen technological advances and perhaps equally important by an assessment of the fundamental ability of human warfighters in today's battlespace. If this goal of better-than-human performance is achieved, even if still imperfect, it can result in a reduction in noncombatant casualties and property damage consistent with adherence to the Laws of War as prescribed in international treaties and conventions, and is thus worth pursuing vigorously.

Keywords: robotics, unmanned systems, autonomy, lethal autonomous systems, ethics, laws of war

1. Introduction

The trend is clear: warfare will continue and autonomous robots will ultimately be deployed in its conduct. Referring to the improving technology of the day and its impact on the inevitability of warfare, Clausewitz stated 'the tendency to destroy the adversary which lies at the bottom of the conception of War is in no way changed or modified through the progress of civilization' (Clausewitz 1832). More recently, Cook observed 'The fact that constraints of just war are routinely overridden is no more a proof of their falsity and irrelevance than the existence of immoral behavior 'refutes' standards of morality: we know the standard, and we also know human beings fall short of that standard with depressing regularity' (Cook 2004). Given this, questions then arise regarding if and how these systems can conform as well or better than our soldiers with respect to adherence to the existing Laws of War. If achieved, this would result in a reduction in collateral damage, i.e., noncombatant casualties and civilian property. The body of research conducted in our laboratory (Arkin 2009, Arkin and Ulam 2009, Arkin et al 2009) focuses on this issue directly from a design perspective. As robots are already faster, stronger, and in certain cases (e.g., chess playing) smarter than humans, is it that difficult to believe they will be able to treat us more humanely in the battlefield than we do each other?

This is no simple task, however. In the fog of war it is hard enough for a human to be able to effectively discriminate whether or not a target is legitimate. Fortunately for a variety of reasons, it may be anticipated, despite the current state of the art, that in the future autonomous robots may be able to perform better than humans under these conditions, for the following reasons:

1. The ability to act conservatively: i.e., they do not need to protect themselves in cases of low certainty of target identification. Autonomous armed robotic vehicles do not need to have self-preservation as a foremost drive, if at all. They can be used in a self-sacrificing manner if needed and appropriate without reservation by a commanding officer. There is no need for a 'shoot first, ask-questions later' approach.

2. The eventual development and use of a broad range of robotic sensors better equipped for battlefield observations than humans currently possess. This includes technological advances in electro-optics, synthetic aperture or wall penetrating radars, acoustics, and seismic sensing, to name but a few.

3. Unmanned robotic systems can be designed without emotions that cloud their judgment or result in anger and frustration with ongoing battlefield events. In addition, 'Fear and hysteria are always latent in combat, often real, and they press us toward fearful measures and criminal behavior' (Walzer 1977). Autonomous agents need not suffer similarly.

4. Avoidance of the human psychological problem of 'scenario fulfillment' is possible, a factor believed partly contributing to the downing of an Iranian Airliner by the USS *Vincennes* in 1988 (Sagan 1991). This phenomenon leads to distortion or neglect of contradictory information in stressful situations, where humans use new incoming information in ways that only fit their pre-existing belief patterns, a form of premature cognitive closure. Robots need not be vulnerable to such patterns of behavior.

5. They can integrate more information from more sources far faster before responding with lethal force than a human possibly could in real-time. This data can arise from multiple remote sensors and intelligence (including human) sources, as part of the Army's network-centric warfare concept (McLouglin 2006) and the concurrent development of the Global Information Grid (DARPA 2007). 'Military systems (including weapons) now on the horizon will be too fast, too small, too numerous and will create an environment too complex for humans to direct' (Adams 2002).

6. When working in a team of combined human soldiers and autonomous systems as an organic asset, they have the potential capability of independently and objectively monitoring ethical behavior in the battlefield by all parties and reporting infractions that might be observed. This presence alone might possibly lead to a reduction in human ethical infractions.

Aside from these ethical considerations, autonomous robotic systems offer numerous other potential operational benefits to the military: faster, cheaper, better mission accomplishment; longer range, greater persistence, longer endurance, higher precision; faster target engagement; and immunity to chemical and biological weapons among

others (Guetlein 2005). All of these can enhance mission effectiveness and serve as drivers for the ongoing deployment of these systems. But our research (Arkin 2009) focuses on enhancing ethical benefits by using these systems, ideally without eroding mission performance when compared to human warfighters.

2. Human Failings in the Battlefield

It is not my belief that an autonomous unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform more ethically than human soldiers are capable of performing. Unfortunately the trends in human behavior in the battlefield regarding adhering to legal and ethical requirements are questionable at best. 'Armies, armed groups, political and religious movements have been killing civilians since time immemorial' (Slim 2008, p. 3). The dangers of abuse of unmanned robotic systems, such as the Predator and Reaper, in war are well documented, which occurs even when a human operator is directly in charge (Sullivan 2010, Filkins 2010, Adams 2010). Battlefield atrocities[1] are as old as warfare. 'Atrocity… is the most repulsive aspect of war, and that which resides within man and permits him to perform these acts is the most repulsive aspect of mankind' (Grossman 1995, p. 229).

Humanity's propensity to wage war has gone unabated for as long as history has been recorded. One could argue that man's greatest failing is being on the battlefield in the first place. Immanuel Kant asserted 'War requires no motivation, but appears to be ingrained in human nature and is even valued as something noble' (Kant 1985, p. 125). Even Albert Einstein, who remained a pacifist well into his 50s, eventually acknowledged 'as long as there will be man, there will be war' (Isaacson 2007, p. 494). Sigmund Freud

was even more to the point: '… there is no likelihood of our being able to suppress humanity's aggressive tendencies' (Isaacson 2007, p. 382). In this article, however, we are concerned for the large part with the shortcomings humanity exhibits during the conduct of war (*jus in bello*) as opposed to what brought us there in the first place (*jus ad bellum*).'The emotional strain of warfare and combat cannot be quantified' (Bourke 1999, p. 232), but at least there has recently been a serious attempt to gather data on that subject. A recent report from the Surgeon General's Office (Surgeon General 2006) assessing the battlefield ethics and mental health of soldiers and marines deployed in Operation Iraqi Freedom is disturbing. The following findings are taken directly from that report:

1. Approximately 10% of Soldiers and Marines report mistreating noncombatants (damaged/destroyed Iraqi property when not necessary or hit/kicked a noncombatant when not necessary). Soldiers that have high levels of anger, experience high levels of combat or those who screened positive for a mental health problem were nearly twice as likely to mistreat noncombatants as those who had low levels of anger or combat or screened negative for a mental health problem.

2. Only 47% of Soldiers and 38% of Marines agreed that noncombatants should be treated with dignity and respect.

3. Well over a third of Soldiers and Marines reported torture should be allowed, whether to save the life of a fellow Soldier or Marine or to obtain important information about insurgents.

4. 17% of Soldiers and Marines agreed or strongly agreed that all noncombatants should be treated as insurgents.

5. Just under 10% of Soldiers and Marines reported that their unit modifies the ROE to accomplish the mission.

6. 45% of Soldiers and 60% of Marines did not agree that they would report a fellow soldier/marine if he had injured or killed an innocent noncombatant.

7. Only 43% of Soldiers and 30% of Marines agreed they would report a unit member for unnecessarily damaging or destroying private property.

8. Less than half of Soldiers and Marines would report a team member for engaging in unethical behavior.

9. A third of Marines and over a quarter of Soldiers did not agree that their NCOs and Officers made it clear not to mistreat noncombatants.

10. Although they reported receiving ethical training, 28% of Soldiers and 31% of Marines reported facing ethical situations in which they did not know how to respond.

11. Soldiers and Marines are more likely to report engaging in the mistreatment of Iraqi noncombatants when they are angry, and are twice as likely to engage in unethical behavior in the battlefield than when they have low levels of anger.

12. Combat experience, particularly losing a team member, was related to an increase in ethical violations.

This formal study, although at the very least disconcerting, is by no means the first report of battlefield atrocities. 'Atrocious behavior was a feature of combat in the two world wars, as well as in Vietnam' (Bourke 1999, p. 163). One sociological study of fighting in Vietnam, pointed out that for all men in heavy combat, 1/3 of men in moderate combat, and 8% in light combat had seen atrocities or committed or abetted noncombatant murder (Strayer and Ellenhorn 1975). These numbers are staggering.

Possible explanations for the persistence of war crimes by combat troops are discussed in (Bill 2000, Parks 1976, Parks 1976a, Danyluk 2000, Slim 2008). These include:

- High friendly losses leading to a tendency to seek revenge.
- High turnover in the chain of command, leading to weakened leadership.
- Dehumanization of the enemy through the use of derogatory names and epithets.

• Poorly trained or inexperienced troops. This lack of training is not just in being a good soldier, but also in understanding the Laws of War.

• No clearly defined enemy.

• The issuance of unclear orders where the intent of the order may be interpreted incorrectly as unlawful.

• Shortage of personnel has been associated in producing stress on combatants that can lead to violations.

• Youth and immaturity of troops.

• An overpowering sense of frustration.

• Pleasure from the power of killing.

• External pressure, e.g., for a need to produce a high body count of the enemy.

There is clear room for improvement, and autonomous systems may help. Bourke points out that modern warfare enables violent acts in ways unlike before. Now, 'Combatants were able to maintain an emotional distance from their victims largely through the application of… technology' (Bourke 1999, p. xvii). This portends ill for the reduction of atrocities by soldiers. We now have bombs being dropped in Afghanistan and Iraq by UAV operators from almost halfway around the world in Nevada (Ure 2008). This use of technology enables a form of 'numbed killing'. Bourke further notes that there is now a 'technological imperative' to make full use of the new equipment provided. Although technological warfare has reduced the overall number of soldiers required to wage war, the price is that technology, while increasing the ability to kill, decreases 'the awareness that dead human beings were the end product'. When killing at a maximum range, one can pretend they are not killing human beings, and thus experience no regret (Grossman 1995). This physical distance detaches the warfighters from the consequences of the use of their weaponry.

The psychological consequences on our servicemen and women in Afghanistan and Iraq have reached record levels. In 2007 alone, 115 soldiers committed suicide, up from 102 the previous year; 24% of the suicides were those on their first deployment, and 43% were those who had returned from deployment. The suicide rates of active duty soldiers as of August 2008 'were on pace to surpass both last year's numbers and the rate of suicide in the general U.S. population for the first time since the Vietnam war, according to U.S. Army officials' (Mount 2008, p.1). This unfortunately was confirmed in July of 2010 (Fifield 10). A statistically significant relationship has been established between the suicide attempts and the number of days spent deployed in Iraq or Afghanistan. To make matters worse, this is coupled with 'a growing number of troops diagnosed with post traumatic stress disorder' (Sevastopulo 2008, p. 1).

These psychiatric casualties are quite significant and common (Grossman 1995): In World War II alone more than 800,000 men were classified unfit due to psychiatric reasons, but an additional 504,000 (approximately fifty divisions) were subsequently rendered unfit as a result of psychiatric collapse after induction; In the 1973 Arab-Israeli war one-third of the Israel casualties were psychiatric in origin, twice the number of dead troops. One WWII study showed that after 60 days of continuous combat, 98% of all surviving troops suffer psychiatric trauma of some sort (Swank and Marchand 1946). These long-term exposures to combat are a recent trend in battle, emerging in the 20th century. The psychiatric damage can result in many forms: battlefield fatigue, conversion hysteria, confusional states, anxiety states, obsession and compulsive states, and character disorders (Grossman 1995). The overall effect on the ability to wage war is obvious, let alone the damage to a nation's surviving citizens.

Creating true warfighters in the first place is a daunting challenge. 'No matter how thorough the training, it still failed to enable most combatants to fight' (Bourke 1999, p. 61). In World War II most men simply did not kill. In one U.S. Army interview of 400 men, only 15% of the men had actually fired at enemy positions (at least once) during an engagement despite the fact that 80% had the opportunity to do so (Marshall 1947). There was no observed correlation between the experience, terrain, nature of the enemy, or accuracy of enemy fire on this percentage.

This applied to both land and air forces. One study of the Korean War indicated that 50% of F-86 pilots never fired their guns and only 10% of those had actually hit a target (Sparks and Neiss 1956). During World War II, most fighter pilots never even tried to shoot anyone down, let alone succeeding. Less than 1% of the pilots accounted for 30-40% of all downed enemy aircraft (Grossman 1995, p. 31).

One conclusion of this is that human soldiers, although not cowardly, lacked an 'offensive spirit'. One possible reason for this lack of aggressiveness centers on the use of long-distance weapons making battlefields 'lonely' and the feeling the enemy was not real but a phantom. This dehumanization of the enemy also quells guilt in killing (Bourke 1999).

The soldiers in the field are not alone in their complicity. 'Atrocities are the dark secret of military culture' (Danyluk 2000, p.38). 'Servicemen of all ranks were unperturbed by most of these acts of lawless killing' (Bourke 1999, p. 173). In Vietnam, combat commanders viewed the Laws of War as 'unnecessary' and 'unrealistic' restraining devices which would decrease the opportunity for victory (Parks 1976, p. 21). A lawyer, defending one General's decision not to initiate a court martial for suspected

war crimes violations, stated ‘It’s a little like the Ten Commandments – they’re there, but no one pays attention to them’ (Hersh 1971, p. 119).

Nonetheless our military aspires to higher ethical performance. General Douglas MacArthur stated:

> The soldier, be he friend or foe, is charged with the protection of the weak and unarmed. It is the very essence and reason for his being. When he violates this sacred trust, he not only profanes the cult, but threatens the very fabric of international society. (Park, 1976, p.18)

In addition, the impact of atrocities on public opinion, as clearly evidenced by the My Lai incident in the Vietnam War, and the consequent effect on troop morale are secondary reasons to ensure that events like these are prevented. Civilians are unfortunately killed during war by other humans for manifold reasons (Slim 2008):

- Genocidal thinking – ethnic or racial cleansing of populations
- Dualistic thinking – dividing host populations into the ‘good guys’ and the ‘bad guys’
- Power dominance and subjugation – power lust and to exert force
- Revenge – emotional striking back for perceived wrongs
- Punishment and forced compliance – to shape the behavior of civilian populations
- Utility – it furthers the war strategically
- Asymmetrical necessity – tactical killing of civilians due to an inferior military position.
- Profit – mercenary and looting activity
- Eradicating potential – pre-emptive removal of civilians who might otherwise become warfighters in the future
- Recklessness – shooting anything that moves, or other forms of indiscriminate killing

• Reluctant killing – through human error or accident, collateral damage.

• Collective and sacrificial thinking – killing of groups rather than individuals, they must be sacrificed for a greater good

These forms of thinking are alien to current artificial intelligence efforts and likely are to remain so. Armed autonomous systems need not nor should be equipped with any of these forms of unacceptable human rationalization or action.

A primary conclusion is that it seems unrealistic to expect normal human beings by their very nature to adhere to the Laws of Warfare when confronted with the horror of the battlefield, even when trained. As a Marine Corps Reserves Captain commented: 'If wars cannot be prevented, steps can be taken to ensure that they are at least fought in as ethical a manner as possible' (Danyluk 2000, p. 38). One could argue that battlefield atrocities, if left unchecked may become progressively worse, with the progression of stand-off weapons and increasing use of technology. Something must be done to restrain the technology itself, above and beyond the human limits of the warfighters themselves. This is the case for the use of ethical autonomy in unmanned systems.

3. A Way Forward

Research in our laboratory has provided the motivation, philosophy, formalisms, representational requirements, architectural design criteria, recommendations, and test scenarios to design and construct an autonomous robotic system architecture capable of the ethical use of lethal force (Arkin 2009). These first steps toward that goal, however, are very preliminary and subject to major revision, but at the very least they can be viewed as the beginnings of an ethical robotic warfighter. The primary goal remains to enforce international humanitarian law (or the Laws of Armed Conflict (LOAC)) in the

battlefield in a manner that is believed achievable, by creating a class of robots that not only comply with the restrictions of international law, but in fact outperform human soldiers in their ethical capacity under comparable circumstances. If successful this will result in the saving of noncombatant life and property, ideally without erosion of mission performance. It is too early to tell whether this venture will be successful. There are daunting problems remaining:

- The transformation of International Protocols and battlefield ethics into machine-usable representations and real-time reasoning capabilities for bounded morality using modal logics.

- Mechanisms to ensure that the design of intelligent behaviors only provide responses within rigorously defined ethical boundaries.

- The development of effective perceptual algorithms capable of superior target discrimination capabilities, especially with regard to combatant-noncombatant status.

- The creation of techniques to permit the adaptation of an ethical constraint set and underlying behavioral control parameters that will ensure moral performance, should those norms be violated in any way, involving reflective and affective processing.

- A means to make responsibility assignment clear and explicit for all concerned parties regarding the deployment of a machine with a lethal potential on its mission.

Hopefully the goals of our limited effort will fuel other scientists' interest to assist in ensuring that the machines that we as roboticists create fit within international and societal expectations and requirements. My personal hope would be that they will never be needed in the present or the future. But mankind's tendency toward war seems overwhelming and inevitable. At the very least, if we can reduce civilian casualties in compliance with applicable protocols of the Geneva Conventions and the ideals

enshrined within the Just War tradition, the result will have constituted a significant humanitarian achievement, even while staring directly at the face of war.

ENDNOTES

[1] Atrocity here is defined as the killing of a noncombatant: either a civilian or a former combatant who has attained *hors de combat* status by virtue of surrender or wound.

References

Adams. J., "US defends unmanned drone attacks after harsh UN Report", *Christian Science Monitor*, June 5, 2010.

Adams, T., "Future Warfare and the Decline of Human Decisionmaking", *Parameters*, U.S. Army War College Quarterly, Winter 2001-02, pp. 57-71.

Arkin, R.C., *Governing Lethal Behavior in Autonomous Systems,* Taylor and Francis, 2009.

Arkin, R.C. and Ulam, P., "An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions", *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR, Dec. 2009.

Arkin, R.C., Wagner, A., and Duncan, B., "Responsibility and Lethality for Unmanned Systems: Ethical Pre-mission Responsibility Advisement", *Proc. 2009 IEEE Workshop on Roboethics*, Kobe JP, May 2009.

Bill, B. (Ed.), *Law of War Workshop Deskbook*, International and Operational Law Department, Judge Advocate General's School, June 2000.

Bourke, J., *An Intimate History of Killing*, Basic Books, 1999.

Clausewitz, C. Von, "On the Art of War", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner 2005), Pearson-Prentice Hall, pp. 115-121, 1832.

Cook, M., *The Moral Warrior: Ethics and Service in the U.S. Military*, State University of New York Press, 2004.

Danyluk, S., "Preventing Atrocities", *Marine Corps Gazette*, Vol. 84, No. 6, pp. 36-38, Jun 2000.

DARPA (Defense Advanced Research Projects Agency) Broad Agency Announcement 07-52, *Scalable Network Monitoring*, Strategic Technology Office, August 2007.

Fifield, A., "U.S. Army Suicide Rate Exceeds National Average", Financial Times, http://www.ft.com/cms/s/0/2c662840-9b74-11df-8239-00144feab49a.html, July 30, 2010, accessed 7/21`/2010,

Filkins, D., "Operators of Drones are Faulted in Afghan Deaths", *New York Times*, May 29, 2010.

Grossman, D., *On Killing: The Psychological Cost of Learning to Kill in War and Society*, Little, Brown and Company, Boston, 1995.

Guetlein, M., "Lethal Autonomous Systems – Ethical and Doctrinal Implications", Naval War College Joint Military Operations Department Paper, February 2005.

Hersh, S., "A Reporter at Large. The Reprimand", *The New Yorker*, October 9, p. 119, via (Bourke 99), 1971.

Isaacson, W., *Einstein: His Life and Universe*, Simon and Schuster, 2007.

Kant, I., *Perpetual Peace and Other Essays on Politics, History, and Morals*, trans. T. Humphrey, Hackett, Indianapolis, 1985.

Marshall, S.L.A., *Men Against Fire: The Problem of Battle Command in Future War*, New York: William Morrow, 1947.

McLoughlin, R., "Fourth Generation Warfare and Network-Centric Warfare", *Marine Corps Gazette*, September 15, 2006, https://feedback.mcamarines.org/gazette/06mcloughlin.asp, accessed 7/30/2010.

Mount, M., "Army Suicide Rate Could Top Nation's This Year", CNN.com, Sept. 9, 2008, http://www.cnn.com/2008/HEALTH/09/09/army.suicides/ (accessed 7/31/10).

Parks, W.H., "Crimes in Hostilities. Part I", *Marine Corps Gazette*, August 1976.

Parks, W.H., "Crimes in Hostilities. Conclusion", Marine Corps Gazette, September 1976a.

Sagan, S., "Rules of Engagement", in *Avoiding War: Problems of Crisis Management* (Ed. A. George), Westview Press, 1991.

Sevastopulo, D., "US Army Suicide Cases at Record 115", *Financial Times*, May 29, 2008, http://us.ft.com/ftgateway/superpage.ft?news_id=fto052920081802392265, accessed 7/30/10.

Slim, H., *Killing Civilians: Method, Madness, and Morality in War*, Columbia University Press, New York, 2008.

Sparks, B.W. and Neiss, O., "Psychiatric Screening of Combat Pilots", *U.S. Armed Forces Medical Journal*, Vol. 4, VII.6, June 1956.

Strayer, R., and Ellenhorn, L., "Vietnam Veterans: A Study Exploring Adjustment Patterns and Attitudes", *Journal of Social Issues*, 33, 4, as reported in (Bourke 99) 1975.

Sullivan, R., "Drone Crew Blamed in Afghan Civilian Deaths", *Associated Press*, May 5, 2010.

Surgeon General's Office, Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07, Final Report, Nov. 17, 2006.

Swank, R., and Marchand, W., "Combat Neuroses: Development of Combat Exhaustion", *Archives of Neurology and Psychology*, Vol. 55, pp. 236-47, 1946.

Ure, L., "Armchair Pilots Striking Afghanistan by Remote Control", CNN.com, July 9, 2008, http://www.cnn.com/2008/TECH/07/09/remote.fighters/index.html, (accessed 7/30/10).

Walzer, M., *Just and Unjust Wars,* 4th Ed., Basic Books, 1977.