Chapter 12
Deception and Virtue in Robotic and Cyber Warfare

John P. Sullins

Abstract Informational warfare is fundamentally about automating the human capacity for deceit and lies. This poses a significant problem in the ethics of informational warfare. If we want to maintain our commitments to just and legal warfare, then how can we build systems based on what would normally be considered unethical behavior in a way that our commitments to social justice are enhanced and not degraded by this endeavor, is there such a thing as a virtuous lie in the context of warfare? Given that no war is ever fully just or ethical. And that navigating the near instantaneous life and death decisions necessitated by modern conflicts fully taxes the moral intuitions of even the best trained and well intentioned war fighters. It follows, that we need accurate analysis on whether or not we can construct informational technologies that can help us make more ethical decisions on the battlefield. In this chapter I will focus on the fact that robots and other artificial agents will need to understand and utilize deception in order to be useful on the virtual and actual battlefield. At the same time, these agents must maintain the virtues required of an informational agent such as the ability to retain the trust of all those who interact with it. To further this analysis it is important to realize that the moral virtues required of an artificial agent are very different from those that are required of a human moral agent. Some of the major differences are that a virtuous artificial agent need only reveal its intentions to legitimate users, and in many situations it is actually morally obliged to keep some data confidential from certain users. In many circumstances cyber warfare systems must resist the attempts of other agents, human or otherwise, to change its programming or stored data. Given the specific virtues we must program into our cyber warfare systems, we will find that while human agents have many other drives and motivations that can complicate issues of trust, we will find that in comparison to human agents, artificial agents are far less complex and morally ambiguous. Thus it is conceivable that artificial agent should be actually more successful at navigating the moral paradox of the virtuous lie often necessitated by military conflict.

J. P. Sullins (✉)
Sonoma State University, Rohnert Park, CA, USA
e-mail: john.sullins@sonoma.edu

L. Floridi, M. Taddeo (eds.), *The Ethics of Information Warfare,*
Law, Governance and Technology Series 14, DOI 10.1007/978-3-319-04135-3_12,

12.1 Deceit in Warfare: Dastardly Behavior or Tactical Brilliance?

The ethics of lies in the context of warfare might seem deeply dependent on context. If we assume a strategic or "realist" framework for our ethical decision making, then from the standpoint of one engaged in a deadly struggle, lies are wrong when you or your allies are the victim but may be correct or even obligatory if you or your allies perpetrate the falsehood and a more just political situation obtains because of it. For instance, under this kind of thinking it was wrong for the Japanese to cloak their attack on Pearl Harbor but right for the US to hide the development and deployment of the atomic weapon, assuming that it was wrong for the Japanese government to have started the conflict with the United States but correct for the US to do everything in its power to end the conflict.

With the notable exception of relativism, most ethical systems are much more circumspect when it comes to the propagation of falsehood. For instance, a strict deontologist would argue against deceit, even when it advanced one's immediately perceived interests even if those interests appear virtuous to the actor. Other systems would allow for very limited forms of deceit, more or less, depending on the situation and or the motives of the active agent.[1] For instance a rule utilitarian might be able to support a rule that allows for one to lie when dealing with hostile agents, especially if that lie might eliminate, impede or damage those agents and result in a situation that maximized the values of the particular utilitarian approach espoused by the moral agent in question be that happiness, human flourishing, or adherence to some set or rule utility.

Here we see the flaccidity of trying to approach this problem with the tools of early modern ethical systems. There is no widespread agreement on whether or not it is permissible for ethical agents to be strategically deceitful when they find themselves in dangerous situations. It just depends on what ethical system you chose, some will allow for it while others will not. Professional philosophers become more or less comfortable with these kinds of systematic impasse and dig their heels in deep and defend their particular flavor of one of these systems to the death. But those outside of philosophy are often deeply troubled by the irreconcilability of the major ethical theories and use this paradox as an indictment the entire project of moral philosophy. The philosopher Eric Dietrich has noticed this fact and has argued that it might be beyond human cognitive capabilities to ever move beyond this deadlock and that it is indicative of deeper flaws in the human ability to undertake the task of philosophy in general (Dietrich 2011a). Interestingly enough, Dietrich is not as pessimistic about the possibility of artificial agents that could move beyond the vexing cognitive limitations of human moral agents and he argues in his essay, "Homo Sapiens 2.0 Why We Should Build the Better Robots of Our Nature," that as humans the one and only truly moral action we can achieve would be to help bring

[1] If one holds the view that there is no truth period, then that certainly ends the discussion. For the sake of having something to say I will not address this possibility in this paper. But as we will see, the strict referential truth-value of a statement may be divorced from its effects on moral agents.

these agents to life and then get out of their way so they can proceed to untangle the moral Gordian knot we have tied around ourselves (Dietrich 2011b). Even if it is possible that future artificial agents might make better moral agents, we are still left with the problem of how to design and program artificial moral agents.

There are a growing number of philosophers engaged in theorizing about the possibility of artificial moral agency and or Machine ethics (see, Anderson and Anderson (eds.) 2011; Lin et al. (eds.) 2011; Sullins (ed.) 2011a; Wallach and Allen (eds.) 2010). But these ideas have yet to be fully expressed in actual technologies. One notable exception to this is the work of the roboticist Ronald Arkin of Georgia Tech. Arkin has been researching technical means of providing some ability for artificial weapons systems to reason on their own about whether or not their actions on the battlefield are remaining in accordance to international standards and laws of conduct in war. As part of that work he has developed the initial designs for an "ethical governor" which is a program that monitors the actions of the weapons system as it autonomously patrols the battlefield and seeks to keep the system from straying outside of programed constraints for the system, much like a governor in a mechanical system keeps that system within safe operating parameters (Arkin 2009a, 2009b, 2010; Arkin et al. 2012). As an example, if the system was engaged with some enemy combatants, the weapons targeting systems would be finding and engaging targets, but this ethical governor would monitor these actions and if the situation changed such that there became too much of a possibility for unacceptable damage to civilians or property, or that the system might need to be constrained due to certain rules of engagement or laws of war that were in effect for this mission, then the ethical governing system would take control of the machine and cease firing (ibid.). This is just one of many conceivable systems but what is most interesting here is that the work is not just theoretical. Arkin and his colleagues are approaching this problem as engineers who are working to develop real systems and products, they see ethics as a kind of technology or at least as something that can be expressed through technology. This move was presaged early last century by the philosopher John Dewey who argued that traditional ethics and morality were incapable of adequately confronting the vexing moral issued raised by the new challenges of a global technological society and he argued that they should be reconstructed as a means for determining new methods for improving value judgments (see, Dewey in Gouinlock (ed.) 1994), and the Dewey scholar Larry Hickman argues that this process can be seen as an instrumental or technological approach to ethics (Hickman 1990). Values are a kind of tool that helps guide conduct and these can be revaluated on the basis of empirical evidence gained while operating under the values in question thus allowing for a kind of moral progress as old values confront insurmountable challenges and are replaced by new ones as was required by the great social changes and conflicts that constituted the era Dewey lived in. In this way one is not appealing to a fixed set of norms or some metaphysical *telos* to make moral judgments but rather a society holds to a developmental set of norms that are always open to revision if they confront a serious challenge that they are unable to otherwise successfully mitigate. This instrumental approach to ethics was further clarified by Mario Bunge who recognized that moral statements were often in the

form of conditionals and that could be transformed into a more precise logic or programed into a kind of information technology he called “Technoethics” (Bunge 1977). We can see this approach to ethics mirrored in Arkin’s work though it seems that this is a coincidence and not by design. In this paper we will focus on this method of approaching ethics and morality as it allows for us to move beyond the meta-ethical road blocks posed by traditional moral systems that may otherwise prevent work on the specific moral issues that confront machine ethics. We can now return to the more focused discussion of the proper role of deception in artificial systems.

12.2 Information Content and Ethics

Being deceitful is wrong. This is the initial moral intuition that comes quickly to mind. No society can tolerate lies about important matters and no individual wants to live in a world where one cannot trust others to be truthful with them. A lie is a known falsehood masquerading as truth and knowing the truth is always better than being deceived by something that is false. In this line of reasoning we have been speaking of deceit in terms of only true and false, right and wrong. Another option would be to deny the claim that any particular statement is exactly either the truth or a lie. In fact after a bit more contemplation we can see that it is possible to be deceitful while only uttering true sentences. For instance, a sentence may be strictly true but through omission can still mislead other moral agents. As an example one might ask a local informant if there are any enemy combatants in a particular area. To which the informant truthfully answers “no” but also knowing full well that they intend to return soon but lets the interrogator continue on into the area as if it were safe. On the other hand, there are instances where statements that on appeal to factual referent are patently false, yet may still lead to morally beneficial situations. I am thinking here of the “Platonic” lie, or statements like, “my love for you is endless.” The former is a paternalistic falsehood that is delivered in an attempt to help the one deceived not suffer unnecessarily or to question something they are incapable of understanding, while the later example is a promise that is improbable in the extreme but a lovely sentiment nonetheless and emotionally very satisfying to the person it is spoken to.

Thus, technically speaking, information content can have a true referent and beneficial consequences, a true referent and maleficent consequences, a false referent and beneficial consequences, and finally a false referent and maleficent consequences (see Fig. 12.1). Therefore, in any give situation an agent must determine the correct amount of truth and falsehood needed to produce a beneficial consequence. It follows then that a moral agent, who wishes to produce beneficial situations, may be called upon to knowingly deceive.

Even this more complex notion of deceit is not entirely adequate. We have so far dodged the question of what makes a particular consequence beneficial or maleficent. Again we seem to be lost in a conceptual muddle as something that seems beneficial to me, such as my gaining access to your savings account, might from

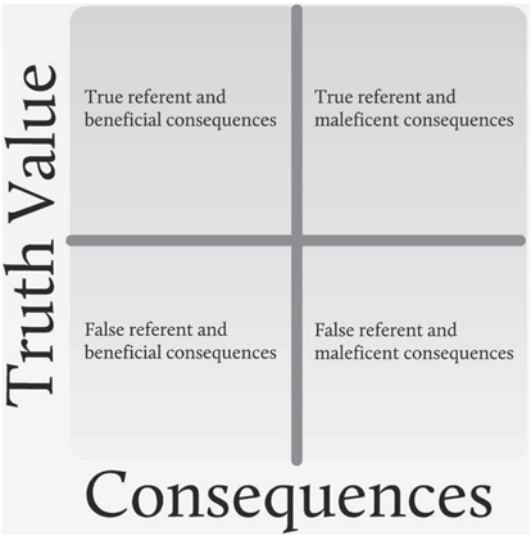

Fig. 12.1 Information content and ethics

your point of view be quite bad. There are, of course, longstanding and venerable arguments in philosophy that attempt to tie this judgment of value to some kind of universal form or numina which can serve as the final arbiter of the beneficial value of some given action. Unfortunately, these titanic debates have yet to resolve into satisfactory answers that are useful for the design of artificial moral agents.

Luckily for us, in the particular context of this discussion we will be able to work around this problem. For now, it is not necessary to attempt to resolve these heady multi-generational debates here as we are engaged in a very well defined arena of discussion. Robotic or cyber warfare (Informational warfare for short), as a subset of warfare in general has a set of very specific set of rules, laws and codes of conduct associated with it which has been developed through international negotiations. On the theoretical side as well a very through philosophical analysis in the form of just war theory has evolved over millennia that serves to inspire moral conduct in the declaration of and execution of war.

Briefly put, just war theory imposes duties on those who would start or fight in wars. Wars can only be propagated by dully constituted authorities whom are motivated by right purpose and only as a last resort to all other means of avoiding conflict (Jus ad Bellum). While warfighters must hold all noncombatants immune to violence while making sure to be proportionate in the violence they can justly impose on enemy combatants and these actions must produce more good than harm (Jus in Bello).

While it would be impossible to argue that these rules, laws and norms are perfect, they are reasonable and provide a good place to ground our discussion. So, for the purposes of this paper then a beneficial situation will be defined as one that does not strongly contradict just war theory or the international rules and law of war.

While the theory of Just War will not be thoroughly questioned in this particular chapter, I do want to make sure I am on the record in advising that they are constantly under discussion and that they are subject to change in light of new evidence and moral challenges. We are in a historical epoch in which the technology of war is changing more radically than it did even under the introduction of gunpowder, so it is obvious that many of our long cherished notions of just warfare will be under stress and will need modification.

Another assumption I would like to make here is that that informational dissolution is nearly always a maleficent outcome for any action and a good measure of whether an action is beneficial or not. Informational dissolution is simply the loss of information, whether that loss comes in the form of the annihilation of a computational system by a virus or worm or the loss of life and memory occasioned by a projectile through a brain, generally speaking both of these are a bad thing. Any action that results in informational dissolution will receive its negative evaluation in direct proportion to the irretrievability of the information lost. For instance, the destruction of the Mona Lisa would be far worse than the destruction of one of a thousand photocopies of the Mona Lisa.

When discussing information in this way it is important to be clear about what is meant by the term "information." Paradoxically, there is no completely satisfying answer to what information is, though the term is obviously very useful nonetheless. Here information is meant is a way that is a bit more philosophically stronger than the way one might define information in an engineering context. Engineers will be happy to define information in the manner of Claude Shannon who describes it as a "signal" which is the ordered set of symbols that can be communicated between two or more agents along some channel with little or no "noise" or loss in the accuracy of the original message (Shannon and Weaver 1949). In addition to this definition of "information" we need to add here a more deeply ontological claim. Information is also something that either constitutes or is very closely correlated with existence itself. This is the basic intuition that motivates the emerging fields of information philosophy and information ethics (see, Floridi 2011). We are straying close to another metaphysical wormhole here as if we take these propositions seriously, then given that everything is constituted of information, it would seem that all warfare is informational warfare. As interesting as that idea is, let's just back away from it for now and return to the more prosaic understanding of information. This way we can see that without overriding moral arguments, informational dissolution caused by robotic or cyber warfare is not a beneficial outcome and we can measure that by the extent to which the information lost is difficult to retrieve or replace.

Finally, there is one more term that needs to be clarified before we can go on. Here we will use the term "virtue" to refer to the proper reasoning, programming, or habits of artificial and/or natural moral agents which are needed to ensure that one's actions bring about beneficial conduct. This should allow us to build an argument that in some cases a virtuous artificial agent could use deception to bring about a beneficial situation measured in terms of avoiding informational dissolution. While this notion of virtue is not precisely the same as is used in ancient or modern virtue

ethics, it is still a position that is defaceable and will be useful for building the following arguments.

12.3 Informational Warfare and the Commitment to Just War

No form of warfare is ever fully just or ethical. This is due to its destructive nature; warfare always creates immediate maleficent outcomes. As the famous American civil war general Tecumseh Sherman observed, "War is all Hell."[2] At best it can only serve as a way to assure that one's enemies are dealt a greater share of that hell than they can deal to you. The destruction of warfare might also be mitigated if the war is just as it is claimed in just war theory that if the reason for the war is just and the war is fought justly and ethically, then the short term evil of the destruction and violence of war can lead to long term good in the form of a stronger and lasting peace.

If we follow this reasoning, then we must conclude that although informational warfare will always contribute to short-term maleficent outcomes in the form of irretrievable loss of life, property and information, but if these war fighting tools are used in the propagation of just war, then it might lead to long term good. This leads us to our first claim; Informational warfare must be committed to the propagation of only just war.

Technologies embody the moral commitments of their makers and users. This means that the design of informational warfare technologies can lead to systems that either enhance our commitment to just war or degrade it. The modern battlefield has evolved into a place where a great deal of information is available to the war fighter, which is good only if that information can be quickly processed and the useful and accurate information sifted from the false and useless. Acting on poor information can lead to unintended damage and casualties. Making quick and accurate decisions that lead to ethical outcomes is a taxing activity that can quickly overwhelm the cognitive capacity of unaided human agents. The job of informational warfare is to assist the war fighter in making good decisions. But it is increasingly the case that informational warfare must be more that simply data acquisition and management tools, due to the pressures to make these decisions in a faster and more efficient manner, it is inevitable that more and more of the processing and synthesis of information as well as decision making based on this information be done by the informational system itself (Singer 2009).

While the situation on the modern battlefield may demand these capabilities from our informational warfare systems, giving them this capability is much easier said than done. It is not my purpose her to outline the many obstacles to the development of these systems. Instead I wish to grant that these problems are only technical

[2] More specifically he is quoted as saying, "There is many a boy here today who looks on war as all glory, but, boys, it is all hell," at a speech given April 11, 1880 in Columbus Ohio.

issues that will be solved sooner or latter. What I want to explore here is the question of which virtues do we need to program into artificial moral agents as they become more autonomous in order to maintain our commitment to just war.

12.4 The Virtues of Informational Systems

When we talk of virtue it is easy to anthropomorphize our machines and miss apply the concept to artificial agents. Ethical systems based on the concept of virtue are all designed with the basic assumption that we are only dealing with other human moral agents. While virtue ethics is a powerful system for understanding and refining human ethical judgment, it must be adapted for use by artificial agents unless and until those agents achieve human level intelligence and become interested in human style eudemonia.

This can be illustrated by looking at Aristotle's famous illustration of virtue, courage. He argues that true courage exists in a mean position between cowardice and foolhardiness. Courage is the willingness to risk harm in the pursuit of protecting other moral agents or important ideals. The exemplar of this would be the virtuous soldier who takes risks to protect other human agents and justice, but does not simply throw his life away in a pointless gesture of bravado. What makes the behavior so exceptional and worthy of praise is that the virtuous soldier may lose her or his life in the process, so they are risking literally everything for altruistic reasons. None of this makes any sense when applied to an artificial agent such as a military robot or cyber warfare system. How can these systems display this kind of courage? They risk very little, they have no sense of existence nor do they have their own goals or desires. More importantly, they do not have beliefs about their own goals and desires which they can modify to become more virtuous in the classical sense. Thus their actions are not entirely their own and cannot be said to be motivated by anything like human courage.

Even if the same action they commit would be considered courageous if done by a human agent. Human medics and corpsmen are noted for their many acts of courage through the centuries saving wounded warfighters often while under enemy fire. Now imagine a cleverly designed and programed robot that rescues a wounded human warfighter under similar enemy fire. Would that machine be worthy of the same kind of commendation we might give a human medic or corpsman? The question here is much more difficult to answer. I have argued that "Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities" (Sullins 2011b). So we might grant the machine moral agency depending on the autonomy, intentionality and responsibility of the machine in question, but it would take quite a lot of these three requirements before we might be tempted to claim that the machine was exhibiting excellence in the virtue of courage.

Of course this all changes if these systems develop, or are given the conscious understanding of their own existence and develop unique personalities that can be

risked. Then these systems might also develop courage. But here we are talking about far future systems and are losing our focus on existent and near future technologies. Instead we must look at the kinds of virtues appropriate for informational systems. While these virtues are not necessarily sufficient for human moral agents, they are actually of some value when we are dealing with the restricted or even nonexistent self-awareness of artificial agents as they exist today.

Even though the long list of human virtues are barely applicable to artificial agents causing them to be seemingly impoverished moral agents, there is a potential benefit that can be leveraged. Human moral agents have many conflicting drives and desires that can complicate their ability to act entirely virtuous in any given situation. Artificial moral agents, at least the simple ones we can imagine in the near future, have a much more restricted list of potential virtues and therefore the complex internal moral conundrums should be rarer for them.

It might seem that the argument so far has concluded that traditional virtue ethics might not have much to add to our discussion, but that is only true if we are fixated on human level virtue. Instead we should shift our focus to virtues that are appropriate for artificial agents designed for informational warfare.

The virtues we are about to discuss are inspired by the "CIA" security triad that has been in use by the computer security community for some time now. The acronym "CIA" refers to: Confidentiality, Integrity, and Availability. These represent the desirable qualities that should be expressed by security systems. Confidentiality is used to insure that only authorized individuals have access to stored information. Integrity represents the ability of a security system to keep tabs on who and how any data is modified. Availability is the system's ability to have the data ready and accessible for legitimate users. This is a very sensible list but there have been many alterations to these basic concepts over the years by various interested parties. For instance the Organization for Economic Co-operation and Development (OECD) in the *OECD Guidelines for the Security of Information Systems and Networks* lists to nine separate principles for security professionals: Awareness of the need for security, Responsibility for secure information, Response to issues in a timely manner, Ethics and respect towards users, Democracy should be upheld, Risk assessment must be through, Security design and implementation in all systems and networks, Security management should reflect the above values and, Reassessment of these systems must be regular (OECD 2002). The security guru Bruce Schneier suggests this list: Privacy, Multilevel Security or secrets within secrets, Anonymity (personal and political), Commercial Anonymity, Medical anonymity, Authentication, Integrity, Audit, and Proactive solutions to threats (Schneier 2000). Taken together we can see some overlap in these lists of principles but some seem to refer to the human operators and users of the systems and some obviously refer only to the systems themselves. Next I would like to disentangle these principles with an eye towards application in informational warfare systems themselves.

It is fair to ask here why insist on using the term virtue when security professionals obviously prefer to speak of principles or rules? The main benefit of working with virtues is that they are understood to be the mean between two extremes. A

virtue is a nuanced approach to moral reasoning rather than an all or nothing step function. We will see how that works out below.

Informational warfare systems need three foundational virtues; security, integrity and accessibility. Informational security is achieved when the system is able to balance the needs of integrity and accessibility demanded by systems and users at differing security levels. Simply put, data stored at a high security levels must be kept free from modification and deletion by lower security users or outside intruders. Integrity is achieved by balancing the needs of accessibility and security of data use by users of various security levels. This means that data use by low security level users or systems must not be allowed to be contaminated high security level data, though the system must be able to profit from low level information that can be verified, e.g. information obtained from an informant of some sort. Accessibility is obtained when the system correctly balances the needs of data use for all levels of systems and users while maintaining security and integrity. This requires that low security level systems and users have access to the information that they are warranted and that all of their data must be made available to higher security level systems and users if needed and where it is appropriate. In addition to this we can only claim a system is accessible when the system or user is able to access information needed precisely when needed it is needed.

These virtues having been abstracted from the civilian security profession have some interesting ethical commitments that may need modification for use in informational warfare situations. This is due to the fact that these systems must maintain security, integrity and accessibility while at the same time working to deny these very same abilities to enemy informational systems. In the civilian setting we can see from the lists of principles above that ethical security professionals have a strong desire to insure that there systems deal honestly with their legitimate users. For instance, they are not designed to give false information to certain users, but rather to simply deny access to protected information.[3] If a user has the proper security level then the system will become fully open and trustworthy. These systems are designed to be trustworthy and honest. Fine virtues indeed but if informational warfare systems adopt only these virtues, than they may be vulnerable. As we found above, deceit and the understanding that other users and systems might be potentially being deceitful to them is a necessary capability of informational warfare systems.

12.5 Robots, Informational Systems and Deceit

Research in building informational systems that intentionally deceive humans or other systems is only just beginning. Of course many forms of spy and malware work by causing the system they infect to think of them as just another benign sys-

[3] Note that as we discussed earlier in the paper, omission can be used to mislead but I do not think that security professionals are necessarily trying to fool their users in this way.

tem with all the proper security clearances. But this is a very minor form of deceit, just a kind of disguise or camouflage.

Ronald Arkin has begun to be experiment in adding deception to the capabilities of robotic weapons systems which received a good deal of media attention. In fact it received so much media attention that Arkin released a statement on the web stating some of his views on the ethical questions raised by this kind of research (Arkin 2011a).

What Arkin and his colleague in this research Dr. Alan Wagner achieved was to program their small autonomous mobile robots in such a way that they each had the ability to develop a model of what the other machine might be "thinking" was true about the toy world they were operating in and then use that information to deceive the other in a simple game of hide and seek (Wagner and Arkin 2009, 2011). These machines had the ability to do things like construct a false "trail" that the other robot would misread to look for the robot in the wrong hiding place (Wagner and Arkin 2011).

> This involves the use of partner modeling or a simplistic view (currently) of theory of mind to enable the robot to (1) assess a situation; (2) recognize whether conflict and dependence exist in that situation between deceiver and mark, which is an indicator of the value of deception; (3) probe the partner (mark) to develop an understanding of their potential actions and perceptions; and (4) then choose an action which induces an incorrect outcome assessment in the partner. (Arkin 2011a)

Perhaps one might want to quibble with Wagner and Arkin as the exact capabilities of the deceitful robots they built. It is obvious that the machines in question are only capable of deceiving one another and would not be very good at a game of hide and seek played against a human or an animal. But in an informational warfare scenario, often the target will be other computational systems so this research shows that deception of this sort is possible.

Arkin comes to some of the same conclusions seen in this paper above regarding the ethical justification for deceit in artificial systems; he agrees that there is no deontological justification but that it might be arguable on consequentialist grounds (ibid.). He does conclude that:

> The point of this paper is not to argue that robotic deception is ethically justifiable or not, but rather to help generate discussion on the subject, and consider its ramifications. As of now there are absolutely no guidelines for researchers in this space, and it indeed may be the case that some should be created or imposed, either from within the robotics community or from external forces. But the time is coming, if left unchecked, you may not be able to believe or trust your own intelligent devices. Is that what we want?. (ibid.)

Another interesting experiment using a simple autonomous robot that served as a referee in a game. The machine used the occasional strategic lie to keep the players interested and the game going. In this experiment it was really the participant's reactions that were being measured and the experimenters reported that:

> Results include the finding that participants were more accepting of lying by our robot than for robots in general. Some participants found the balancing strategy favorable after being debriefed, while others showed less interest due to a perceived level of unfairness. (Vázquez et al. 2011)

Sharkey and Sharkey (2011b) in an article for the IEEE Robots and Automation Magazine, describe how some forms of deception might be useful in the deployment of carebots for the elderly. We must note that this endorsement is only for certain situations that can clearly benefit the patient.

Peter A. Hancock et al. (2011), of the US Army Research Labs have done a nice literature review of the factors that are effecting users trust of robotic systems in military contexts. The findings of use to us here are that military robots form an integral part of human machine teams and are used to help mitigate the cognitive overload of warfighters attempting to determine accurate situational awareness in combat situations. They also find that trust is a complex human psychological state and that humans in these human-machine teams can place both too much and too little trust in their robotic assets. They also have determined that human, environmental and robot characteristics all impact the level of trust placed in the robot and negative trust can be mitigated by proper training and design (Hancock et al. 2011).

From these initial results it would seem that except for the Army Research Labs report, there is some hesitant support for allowing informational warfare systems to be engaged in some forms of deceit.

In order for that deceit to be ethical, it must be done in such a way that the resulting situation is more beneficial than would obtain had the deceit not occurred. In this context that would require at the minimum that the deceit results in a situation that advances the dictates of the rules of engagement, laws of war and principles of just war that are attendant to the conflict at hand.

As informational warfare systems become more autonomous, they must then be designed with a commitment to the foundational virtues of security, integrity and accessibility. Strategic deceit does not run counter to these virtues but in fact can help maintain them in certain situations.

The main problem we have to worry about here is that building deceit into our systems will violate the cherished notion that computers never lie. Trust and robotics has a troubled relationship (Coeckelbergh 2012). We can see from the Army Research Labs report that there are occasions already where humans working in close partnership with machines on the battlefield distrust the information they are receiving from them. If the machine was known to have the ability to deceive, then this might exacerbate the situation and make the partnership unworkable. For this reason it would be best to design the machines to error on the side of disclosure to legitimate users and only use deceit in the face of enemy threats or in actions to defeat enemy informational warfare systems.

12.6 Conclusions

This paper has shown that ethics can be profitably seen as a kind of technological undertaking designed to test and validate social values. Here we have taken on the task of validating our intuitions on the use of deceit by informational warfare systems. We found that in certain situations (but not all) deceit may be the more

ethical choice in bringing about situations that fulfill our commitment to just war and the rules and laws of war. To achieve this goal we found that informational warfare systems must maintain commitments to the foundational virtues required of an informational agent; security, integrity, and accessibility. These virtues would be insufficient for a human agent but are adequate for the limited artificial agents under discussion here. Systems built with these values in mind will have the ability to retain the trust of all appropriate users who interact with the system. We also found that the virtues of an informational agent are very different from those of a human agent. A virtuous informational agent that is balancing the needs for security, integrity and accessibility needs to reveal its intentions only to its legitimate users while keeping certain bits of data confidential from low security level users, and resist the attempts of intruders into the system and other low security agents whom might wish to change its programming or stored data.

Critics of this position have been worried that any system built with these capabilities might move beyond the control of the human agents deploying it or might even be cynically used by humans in a way to deny responsibility for any harm committed by the informational warfare system. One should not deeply worry about the responsibility gap for the commitment of war crimes as argued by Robert Sparrow (see, Sparrow 2007). It would be unrealistic to let the owners and operators of some informational warfare machine off the hook due to the autonomy of the systems deployed. This issue is addressed nicely by a workgroup from the US National Science Foundation and their findings are summed up in a document informally known as “The Rules” and rule 1 clearly states that: The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact.

Finally, while the virtuous lie is an old idea first formally argued for by Plato in The Republic. It has always been a very fraught move morally. One always has to ask if they are telling the lie for the good of the other or simply as an expedient for themselves. It is very easy to fool oneself into thinking their lie is virtuous while those told to them are villainous. But an artificial agent should be actually more successful at navigating the moral paradox of the virtuous lie, given that it has no real stake in the game, no actual wants needs or desires. It is more likely to avoid motivational conflicts common in human agents.

Bibliography

Anderson, M., and S. L. Anderson, eds. 2011. *Machine ethics*. Cambridge: Cambridge University Press.

Arkin, R. 2009a. *Governing lethal behavior in autonomous robots*. New York: Chapman and Hall Imprint, Taylor and Francis Group.

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

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

Arkin, R. C. 2011a. The ethics of robotic deception. http://www.cc.gatech.edu/ai/robot-lab/online-publications/deception-final.pdf. Accessed July 2012.
Arkin, R. C. 2011b. Viewpoint: Military robotics and the robotics community's responsibility. *Industrial Robotics* 38(5).
Arkin, R. C., P. Ulam, and A. R. Wagner. 2012 Mar. Moral decision-making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE* 100(3):571–589.
Arquilla, J. 2010. Conflict, security and computer ethics in Floridi 2010.
Aycock, J., and J. Sullins. 2010. Ethical proactive threat research. Workshop on Ethics in Computer Security Research (LNCS 6054), pp 231–239. New York: Springer.
Bunge, M. 1977. Towards a technoethics. The Monist, 60, 96–107.
Cisco Systems Inc. 2011. Cisco 2011 Annual Security Report: Highlighting global security threats and trends. San Jose: Cisco Systems Inc.
Coeckelbergh, M. 2012. Can we trust robots? *Ethics and Information Technology* 14:53–60.
Crnkovic, G. D., Çürüklü, B. 2012. Robots: Ethical by design. *Ethics and Information Technology* 14:61–71.
Denning, D. 2008. The ethics of cyber conflict. In *The handbook of information and computer ethics*. 1st ed, ed. K. E. Himma, and H. T. Tavanni. Wiley-Interscience.
Dietrich, E. 2011a. There is no progress in philosophy. *Essays in Philosophy* 12 (2).
Dietrich, E. 2011b. Homo Sapiens 2.0 why we should build the better robots of our nature. In *Machine ethics,* ed. M. Anderson, and S. Anderson. Cambridge: Cambridge University Press.
Floridi, L., ed. 2010. *The Cambridge handbook of information and computer ethics*. Cambridge: Cambridge University Press.
Floridi, L. 2011. *The philosophy of information*. Oxford: Oxford University Press.
Gouinlock, J. 1994. *The moral writings of John Dewey*. Buffalo: Prometheus.
Hancock, P. A., D. R. Billings, K. E. Oleson, J. Y. C. Chen, E. De Visser, R. Parasuraman. 2011. A meta-analysis of factors influencing the development of Human-Robot Trust, Army Research Laboratory, ARL-TR-5857, December 2011. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA556734. Accessed July 2012.
Hickman, L. A. 1990. *John Dewey's pragmatic technology*. Bloomington: Indiana University Press.
Himma, K. E., ed. 2007. *Internet security, hacking, counterhacking, and society*. Sudbury Massachusetts: Jones and Bartlett.
Himma, K. E., and H. T. Tavanni, eds. 2008. *The handbook of information and computer ethics*. Wiley-Interscience. 1st edition.
Kaspersky Lab. 2011. *Cyberthreat forecast for 2012*. Moscow: Kaspersky Lab ZAO (available online).
Lin, P., G. Bekey, and K. Abney. 2008. *Autonomous military robotics: Risk, ethics, and design*. Washington, DC: US Department of the Navy, Office of Naval Research (available online).
Lin, P., K. Abney, and G. Bekey. 2011. *Robot ethics: The ethical and social implications of robotics*. Cambridge: MIT Press.
Lovely, E. 2010 Mar 5. Cyberattacks explode in Congress. *Politico* (available online).
Marchant, G., B. Allenby, R. Arkin, E. Barrett, J. Borenstein, L. Gaudet, O. Kittrie, P. Lin, G. Lucas, R. O'Meara, and J. Silberman. 2011. International governance of autonomous military robots. *Columbia Science and Technology Law Review* XII:272–315.
Miller, K. W. 2011. Moral responsibility for computing artifacts: The rules. *IT Professional* 13(3):57–59.
OECD. 2002. OECD Guidelines for the Security of Information Systems and Networks: Towards a Culture of Security, Organisation for Economic Co-operation and Development, Organisation De Coopération Et De Développement Économiques. http://www.oecd.org/dataoecd/16/22/15582260.pdf. Accessed July 2012.
Ramaswamy, S., and H. Joshi. 2009. *Automation and Ethics. Springer Handbook of Automation*. Springer.

Shannon, C. E., and W. Weaver. 1949. *The mathematical theory of communication*. University of Illinois Press.

Sharkey, N. E. (2007 Nov). Automated killers and the computing profession. *IEEE Compute* 40 (11):106–108

Sharkey, N. E. (2008a). Grounds for discrimination: Autonomous robot weapons. *RUSI Defence Systems* 11 (2):86–89.

Sharkey, N. E. (2008b). Cassandra or the false prophet of doom: AI robots and war. *IEEE Intelligent Systems* 23 (4):14–17 (JulyAugust Issue).

Sharkey, N. E. (2009a). Death strikes from the sky: The calculus of proportionality. *IEEE Science and Society* 28:16–19.

Sharkey, N. E. (2009b). Weapons of indiscriminate lethality. *FIfF Kommunikation* 1:26–28

Sharkey, N. E. (2010). Saying "No!" to lethal autonomous targeting. *Journal of Military Ethics* 9 (4):299–313.

Sharkey, N. E. (2011a). Killing made easy: From joystics to politics. In *Robot ethics: The ethical and social implications of robotics,* eds. Lin Patrick, George Bekey, and Keith Abney. Cambridge: MIT Press

Sharkey, A. J. C., and N. E. Sharkey. 2011b. Anthropomorphism and deception in robot care and companionship. *IEEE RAM* 18 (1):32–38.

Schneier, B. 2000. *Secrets and lies: Digital security in a networked world*. New York: Weiley.

Singer, P. W. 2009 *Wired for war: The robotics revolution and conflict in the 21st century*. New York: Penguin.

Sparrow, R. 2007. Killer robots. *Journal of Applied Philosophy* 24 (1):62–77.

Sullins, J. P. 2009. Telerobotic weapons systems and the ethical conduct of war. *APA Newsletter on Philosophy and Computers* 8 (2):21 (P. Boltuc, ed.).

Sullins, J. P. 2011a. Robotics: War and peace. *Philosophy and Technology* 24(3)September.

Sullins, J. P., ed. 2011b. When is a robot a moral agent? In Machine ethics, ed. M. Anderson, and S. L. Anderson.

Tavani, H. T. 2007. The conceptual and moral landscape of computer security. In *Internet security, hacking, counterhacking, and society,* ed. K. E. Himma, 29–45. Sudbury Massachusetts: Jones and Bartlett.

Vázquez, M., A. May, A. Steinfeld. (2011). ShakeTime! A deceptive robot referee. Copyright is held by the author/owner(s), HRI'11, March 6–9, 2011, Lausanne, Switzerland, ACM 978–1-4503-0561-7/11/03. http://delivery.acm.org/10.1145/1960000/1957803/p403-vazquez.pdf?ip=130.157.156.155&acc=ACTIVE%20SERVICE & CFID=97012119 & CFTOKEN=57280102&__acm__=1342826463_29deff0d3692bce36f2a3436192d93e8. Accessed July 2012.

Wagner, A., and R. C. Arkin. (2009). Robot deception: Recognizing when a robot should deceive. *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR.

Wagner, A. R., and R. C. Arkin. 2011. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics* 3 (1):5–26.

Wallach, W., and C. Allen. 2010. *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.