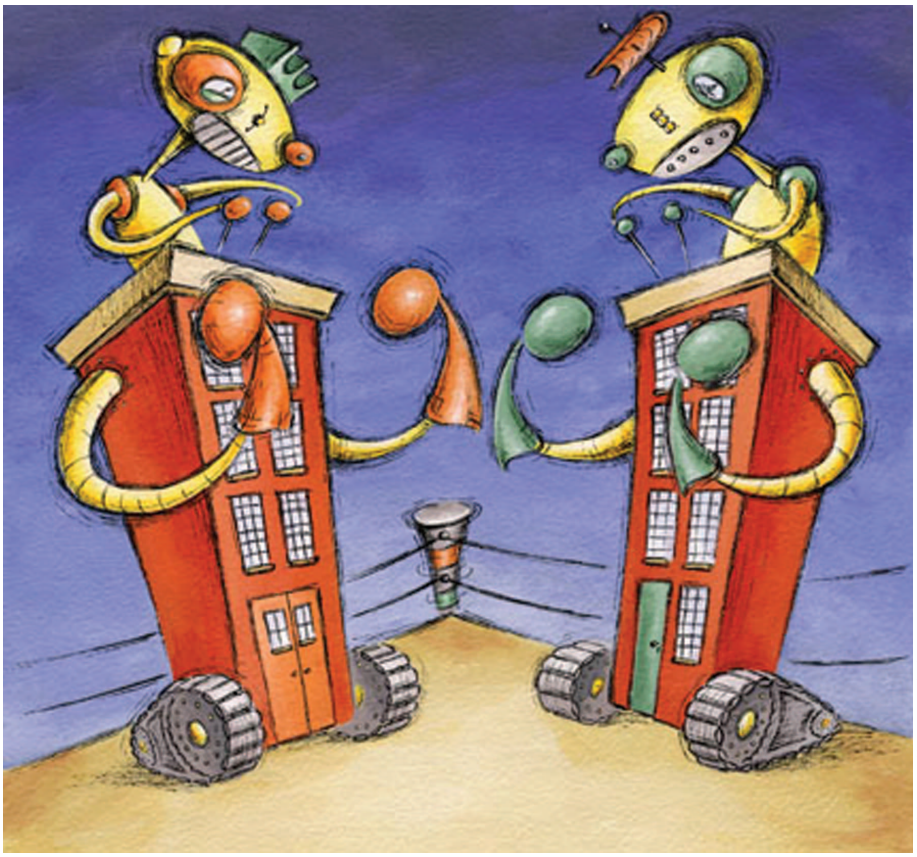
Modeling the Moral User

PETER M. ASARO



There is growing concern about and interest in the ethical design and regulation of autonomous lethal robotics [1], [2], [4], [6], [19]. While these concerns are justified, and their implications for the design of autonomous systems important, it is not merely the autonomous nature of lethal robotics that raises ethical issues in design. Indeed, the design of any safety-critical system that relies upon human decision making, and thus the design of any tele-operated weapons system, has significant implications for the ethical decision making of users. As such, the design of that system should raise ethical considerations for engineers. This is especially true in the design of any weapons system that utilizes sophisticated information processing, data representation, graphical displays, and user controls to manage the use of lethal force.

The fields of human factors, ergonomics, and human-computer interface design have long been engaged in analyzing the implications of various design choices on human performance. More recently, Cummings [9], [10] has argued that ethical decision making is an aspect of human performance that should receive attention in the design of weapons systems interfaces. I believe her argument should be taken further, and I argue that the best way to improve the ethical design of interfaces will be to explicitly model the moral decision-making processes of users.

Digital Object Identifier 10.1109/MTS.2009.931863

I consider here three approaches to designing lethal tele-operated systems—two have been presented in the literature, while the third is new. I will evaluate the strengths and weaknesses of each. The first approach is implied, if not explicit, in Arkin's proposals for the ethical regulation of autonomous systems through a combination of rule-based restrictions and advisory systems (RBAS) [1], [2]. While this approach is meant to apply to a range of different levels of autonomy of the systems it manages, including full autonomy, this article will consider its potential application to tele-operated systems. The second approach is Cummings' [11], which she describes as Value-Sensitive Design (VSD). In VSD, designers consider the implications of particular designs in light of a set of abstract values, which may at times be in conflict. A new, third approach is proposed, following the methods of User-Centered Design (UCD). It endorses modeling the users of these systems, and then using these models to motivate design decisions. I hope to demonstrate just how different these approaches are in their conceptualizations of what constitutes an "ethical" design.

Rule-Based and Advisory Systems (RBAS)

In developing his proposed architecture for the ethical regulation of lethal systems, Arkin [1] draws heavily upon the institutional structures within the military for regulating the conduct of human soldiers in combat. These institutional structures consist of two distinct but related sets of rules: the Laws Of Armed Conflict (LOAC), and the Rules Of Engagement (ROE), as well as Just War Theory (JWT). These rules provide two general ethical principles for lethal decision making: discrimination and proportionality.

Arkin largely treats these sets of rules as given, and assumes that they are 1) internally consistent, 2) that the LOAC, ROE, and JWT are at least mutually compatible, and that following the LOAC and ROE is a practical way of ensuring that the principles of JWT are observed, 3) that the explicit and implicit rules contained within the LOAC, ROE, and JWT can be translated into rules that could be applied within the control architecture of a robotic system, and 4) that issues of ethics and values have largely been settled through the adoption of LOAC, ROE, and the principles of JWT—what is left is a technical matter of ensuring that robotic systems are capable of correctly and reliably carrying them out.

There are problems with each of these assumptions.

While often referred to as if they were a straightforward collection of rules, the LOAC are really a menagerie of international laws and agreements (such as the Geneva Conventions), treaties (such as the Ottawa Treaty, a.k.a. the anti-personnel land-mine ban), and domestic laws regulating the procurement, design, and use of various weapons and tactics. By the very nature of law, these rules are open to challenges and interpretations in various courts, and may not be effectively enforceable. Indeed, in the case of the legal restrictions on torture we have witnessed the issuing of legal interpretations deliberately aimed at shifting military policy and conduct through redefining what acts constitute torture, and who is protected as a "combatant" under the Geneva Conventions. This has led to much confusion amongst soldiers responsible for interrogating prisoners, and would surely lead to a similar confusion among engineers if they tried to design an automated system to decide which actions are prohibited in an interrogation and which are not. Moreover, even when the interpretations of the laws are agreed upon, their application to particular cases can still be contested, and any interpretation depends heavily upon awareness of particular situations.

The ROE are devised to instruct soldiers in specific situations, and take into account not only legal restrictions, but also political, public relations, and strategic military concerns. Thus, it may be imperative not to cross a border into a sovereign territory, or not to fire a weapon until fired upon. Or it might be imperative to avoid damaging certain cultural or religious sites in order not to anger the local population. Or it might be allowed to fire a weapon at vehicles that get too close to a convoy. These rules are devised by military lawyers to suit the needs of specific operations and missions. They often appear ambiguous or vague to the soldiers on the ground who observe situations that do not always fall neatly into the distinctions made by lawyers. Indeed, confusion about the ROE, due to their vagueness or ambiguity, has led to a significant number of civilian causalities in Iraq [15].

Just War Theory is often presented as a consistent, settled theory, though in its canonical formulation by Walzer [21], it is in fact a heterogeneous set of principles, rules of thumb, and values [16]. Moreover, there is a critical ongoing debate about the validity of one of JWT's central tenets, namely the distinction between *jus ad bellum* and *jus en bello*—that is, whether killing can be justified for those fighting an unjust war. Even if this debate were settled in Walzer's favor, the principle of proportionality is abstract, not easily quantified, and highly relative to specific contexts and subjective estimates of value. All of these characteristics make it doubtful that such a principle could be easily built into an automated decision system without over-simplifying it. Indeed, those LOAC which are most often pointed to as examples of proportionality, such as laws against blinding lasers, actually fail the principle of proportionality because they favor

killing over maiming (which is in the interest of nation-states who must provide care for the wounded), whereas in most cases the principle of proportionality actually recommends maiming over killing (for most individuals it is better to be blind than dead).

In summary, I contend that the rules in the LOAC, ROE, and JWT cannot be easily realized within an automated system because they are actually a hodgepodge of laws, rules, heuristics, and principles, all subject to interpretation and value judgments. These rules do have an important role in regulating the conduct and policies of individuals and institutions, mostly because they require people to think about the ethical implications of their actions in certain ways, rather than dictating to them a specific action in a specific situation. If I am correct about this point, then automating these rules would actually undermine the role they play in regulating ethical conduct. It would also explain why designers have sought to keep humans-in-the-loop for the purposes of disambiguation and moral evaluation. As Sir Brian Burridge, commander of the British Royal Air Force in Iraq from 2003 to 2005 puts it:

> Under the law of armed conflict, there remains the requirement to assess proportionality and within this, there is an expectation that the human at the end of the delivery chain makes the last assessment by evaluating the situation using rational judgment. Post-modern conflicts confront us … with ambiguous nonlinear battlespaces. And thus, we cannot take the human, the commander, the analyst, those who wrestle with ambiguity, out of the loop. The debate about the human-in-the-loop goes wider than that [5].

The logical engineering response to this is to focus efforts on advisory or recommendation systems that assist humans in making ethical decisions [2]. Unfortunately, such systems are not without their own ethical hazards.

Value-Sensitive Design (VSD): Weighing Values

Cummings [9]-[10] is clear that the design of user interfaces has significant implications on the moral choices of those who use them. Her principle concern is with systems that provide advice to users, or which will act automatically unless overridden by users. In each case, there is a danger that users will abdicate their responsibility to the automated system. She calls this "automation bias," and defines it as the tendency to trust an automated system, in spite of evidence that the system is unreliable, or wrong in a particular case.

Another ethical problem endemic to tele-operated weapons systems is the creation of "moral buffers," which put psychological distance between users and their actions. This distance diminishes the effects of emotions such as empathy, and reduces the emotional impact of the consequences of one's actions. Cummings uses the example of the Milgram's [17] authority experiments to demonstrate that moving someone out of sight greatly increases the willingness of people to inflict severe pain upon them. It also explains why soldiers are generally more willing to use lethal force as the distance from the people they are killing increases [14].

To avoid both automation bias and the creation of moral buffers, Cummings [10] suggests following the design methods of VSD, which considers the impact of various design proposals on a set of values. Cummings [11] describes a project intended to teach students how to apply VSD to the design of a missile targeting advisory system. After first involving them in a discussion about the relevant moral values in play in the final use of the system, it then challenges them to evaluate their design proposals in relation to these values. This includes empirical testing of interface designs based on how well users do on tasks when the system gives bad advice. What they find is that the designs do influence the choices made by users, and this can directly influence the missile targeting task, and thus the lives of civilians near those targets. The lesson is that making an automated system more efficient (an engineering value) can be at odds with the safety of civilians (a human value).

While this is an excellent tool for teaching engineering ethics through a hands-on design project, I am not convinced this is the best way to design actual systems. While it does endorse an empirical evaluation of user's ethical performances, it does not seek to explain why various designs have the influences they do in any systematic way. Because of this, it cannot provide an overarching design approach or strategy; it can only evaluate given proposals based on a given set of values. There is also the issue of whether the set of values that are generated are comparable, complete, and appropriate, as well as whether they can be properly weighed against each other by engineers within a design process—which in practice is fraught with organizational and economic pressures. Again, while it has enormous value as an ethical exercise, it can only guarantee to demonstrate that one proposed design element is better than another according to a set of abstract values.

What is needed is to increase our empirical knowledge about what kinds of information people use to make various sorts of ethical decisions, how they process that information, and how the presentation and representation of that information influences their performance in ethical decision-making tasks. This

kind of analysis has been done for various tasks, and was the founding problem of the field of human factors during World War II—the use of psycho-physical evaluation of targeting tasks [7], [8], [12].

But these methods have always been applied to situations in which there is a correct performance. What happens if we apply them to cases requiring ethical deliberation and value judgments? To understand how users perform the task of ethical decision making, and thus provide them with the information they need in the form that they need, I propose that we start by modeling ethical decision makers.

User-Centered Design: Modeling the User

User Centered Design (UCD) is a design strategy based on empirical observations of how users actually perform tasks, and using a task model to design interfaces and systems. While empirical studies may have a place in testing RBAS and VSD systems, they are only used to evaluate proposed designs, and do not guide the design methodology directly. I believe that starting with the empirical approach of UCD could improve the ethical design of systems. If one wishes to design a system to perform a task, then it is best to first understand that task. If that task involves humans that are required to make ethical decisions based on interactions with a system, then one should try to understand the nature of the ethical decision-making task from an information-processing perspective. There are two ways to look at this: either as modeling the task environment, or as modeling the user [3]. Thus, I would like to call this approach to ethical design "modeling the moral user." If we want to design a system that enhances the moral performance of users, we should develop a more sophisticated model of a moral user.

These arguments are the same ones put forward for user-centered design more generally [3], [13], [20]. Here they are simply being recast in a moral framework. A typical UCD methodology, following the ISO standard 13407, involves four phases: analysis, design, implementation, and deployment. In each phase one should evaluate the needs and tasks of the user, but it is the analysis phase in which we should be engaging in a critical study of the nature of ethical decision making, and modeling the moral user.

Modeling the moral user will involve three elements. The first element will be to draw upon the methods of cognitive psychology to understand the representations, decision rules, and perceptual and emotional requirements for effective ethical decision making. This might include detailed studies of the psychological roles of authority, propaganda and indoctrination, empathy, sympathy, stress, guilt, vengeance, anger, aggression, and fear.

Second, to draw upon recent work in experimental philosophy into the nature of moral intuition, value comparisons and judgments, and experimental economics into the nature of risk assessment and probability estimation. It will also be crucial to understand to what extent people actually conform to rational standards in ethical decision making, or whether they even formulate ethical decisions in a rational framework.

Third, we will need to consider what we, as a society, wish the ethical standards of the soldiers who use these systems to be. To what extent can we enforce these standards on the soldiers through the technology, and to what extent should they be exposed to physical and psychological risk or harms?

Modeling the Moral User

While I endorse a UCD approach, I am still not convinced that it will be successful in making lethal tele-operated systems ethical. This may be impossible to achieve. However, following the RBAS and VSD approaches, we would have no way of knowing whether the systems which result are in fact ethical, except by analyzing the consequences of fielding such weapons. At least with UCD we might understand what the ethical limitations of such systems would be before they are fielded. Specifically, by developing a more sophisticated model of the moral user, we might recognize various psychological contradictions in the task, as well as begin to understand the range and diversity of ethical considerations that users actually employ in these tasks.

I believe it is an open question whether there is a single model that can effectively capture the range of human moral reasoning. While philosophers have sought theories that might offer consistent moral reasoning, there is little evidence that people actually reason that way, apart from philosophers. It would not surprise me if our empirical studies demonstrated that there are significant variations between cultures, as well as between individuals within a culture. Just as we see variations in kinds and degrees of intelligence and learning styles, we are likely to see variations in ethical reasoning. Indeed, the fact that we recognize differences in the moral character of individuals may simply be due to our recognition of their different values and standards of moral reasoning.

Given that such differences exist, how then ought we design interfaces? Should we design for a moral ideal which no one actually uses? Or for the lowest common denominator? For some hypothetical average? Is it ethical for engineers to impose an ethical model on users who have arrived at a different ethical model through a deliberate and rational process of their own? Or should engineers design interfaces customizable to different ethical styles?

There are further issues regarding the ethical and psychological

impact of these systems on the users. Sparrow [19] expresses a concern about the stress that users of these systems may experience. While on the one hand, they would presumably experience less stress than they would if they were actually physically present in combat, on the other hand these systems also introduce new kinds of stress. One kind is the stress of the dissociation between the world they occupy through the tele-operated robotic system, and the world they enter when they leave the control room. Currently, there are UAV pilots flying lethal missions in Iraq and Afghanistan from control rooms in Nevada, and who report psychological distress at the contrast between their work life and home life. Sparrow also argues that because tele-operated systems do not afford all the actions that human bodies do, they do not allow users to administer first aid or other forms of assistance to wounded soldiers and civilians. Observing tragic events in real-time without being able to intervene can also be emotionally stressful. But again, we can better understand the impact of these stresses by developing a sophisticated empirical model of moral users.

We have only begun to study ethical decision making using the methods of cognitive psychology. Because we know little about individual moral psychology, and how it is influenced by the presentation of information, it seems wise to proceed with an experimental approach that draws upon what is known, and seeks to discover more about how people make ethical choices. What we do know from studying soldiers in combat is that "moral buffers" also provide protection against the psychological stress of killing, and thus post-traumatic stress disorders.

It might turn out that designing a system to enhance the moral awareness of the users of a weapon also increases the chances that they will suffer post-traumatic stress. That is, empirical research may demonstrate that for users to make effective ethical decisions, they need a sufficient amount and kind of information to trigger empathetic and sympathetic emotions. I suspect that these are probably the same information channels that lead to stress and ultimately to post-traumatic stress disorders. If so, then it might be necessary to impose high levels of psychological stress on users in order to improve their ethical performance.

Complex Issues Crucial to Designing Ethical Systems

These are complex issues of both morality and engineering, but they are crucial to designing ethical tele-operated weapons systems, and should not be avoided by our design methodologies. I believe it will be valuable to study these kinds of trade-offs through an effort to model the moral user, even if it turns out that we discover intrinsic paradoxes in the design of ethical weapons systems. And as we learn more about ethical decision making "in the wild" we may also refine our notion of what makes for an ethical design. Thus, while a UCD approach may not ultimately succeed in building more ethical interfaces, it will surely increase our understanding and appreciation of the complexity of human ethical decision making, especially in contexts of war and killing. It is for these reasons that I believe modeling moral users is the best approach to the ethical design of interfaces for tele-operation.

Author Information

Peter Asaro is affliliated with the Center for Cultural Analysis at Rutgers University, and is a member of the Faculty of the Department of Media Studies and Film, The New School University, New York, NY; peterasaro@sbcglobal.net; www.cybersophe.org.

References

[1] R.C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," Georgia Institute of Technology, Technical Report GIT-GVU-07-11, 2007.

[2] R.C. Arkin, A.R. Wagner, and B. Duncan, "Responsibility and lethality for unmanned systems: Ethical pre-mission responsibility advisement," in *Proc. HRI '09* (San Diego, CA), Mar. 11-13, 2009, to be published.

[3] P. Asaro, "Transforming society by transforming technology: The science and politics of participatory design,"*Accounting, Management & Information Technologies*, vol. 10, no. 4, pp. 257-290, 2000.

[4] P. Asaro, "How just could a robot war be?," in *Current Issues in Computing And Philosophy*, P. Brey, A. Briggle. and K. Waelbers, Eds. Amsterdam, Netherlands: IOS, pp. 50-64, 2008.

[5] B. Burridge, "UAVs and the dawn of post-modern warfare: A perspective on recent operations," *RUSI J.*, vol. 148, no. 5, pp. 18-23, Oct. 2003.

[6] J.S. Canning, "A concept of operations for armed autonomous systems," presented at *3rd Ann. Disruptive Technology Conf.* (Washington, DC), Sept. 6-7, 2006.

[7] K.J.W. Craik, "Theory of the human operator in control systems I: The operator as an engineering system," *Brit. J. Psychology*, vol. 38, no. 2, pp. 56-61, 1947.

[8] K.J.W. Craik, "Theory of the human operator in control systems II: Man as an element in a control system," *Brit. J. Psychology*, vol. 38, no. 3, pp. 142-148, 1948.

[9] M.L. Cummings, "Creating moral buffers in weapon control interface design," *IEEE Technology& Society Mag.*, vol. 23, no. 3, pp. 28-33, 41, 2004.

[10] M.L. Cummings, "Automation and accountability in decision support system interface design," *J. Technology Studies*, vol. 32, no. 1, pp. 23-31, 2006.

[11] M.L. Cummings, "Integrating ethics in design through the value-sensitive design approach," *Science & Engineering Ethics*, vol. 12, no. 4, pp. 701-715, Oct. 2006.

[12] P. Galison, "The ontology of the enemy: Norbert Wiener and the cybernetic vision," *Critical Inquiry*, vol. 21, pp. 228-266, 1994.

[13] J. Greenbaum and M. Kyng, *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, NJ: Erlbaum, 1991.

[14] D. Grossman, *On Killing.* Boston, MA: Little Brown, 1995.

[15] C. Hedges and L. Al-Arian, "The other war: Iraq Vets Bear Witness," *Nation*, July 30, 2007.

[16] J. McMahan, "The sources of just war principles," *J. Military Ethics*, vol. 6. no. 2, pp. 91-106, 2007.

[17] S. Milgram, *Obedience to Authority.* New York, NY: Harper and Row, 1975.

[18] R. Sparrow, "Killer robots," *J. Applied Philosophy*, vol. 24, no. 1, pp. 62-77, 2007.

[19] R. Sparrow, "Building a better warbot: Ethical issues in the design of unmanned systems for military applications," *Science and Engineering Ethics,* Dec. 2, 2008.

[20] L. Suchman, "Making work visible," *Commun. ACM*, vol. 38, no. 9, pp. 56-64, 1995.

[21] M. Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York, NY: Basic, 1977.