Toward Modeling and Automating Ethical Decision Making: Design, Implementation, Limitations, and Responsibilities

Gregory S. Reed and Nicholaos Jones

Abstract

One recent priority of the U.S. government is developing autonomous robotic systems. The U.S. Army has funded research to design a metric of evil to support military commanders with ethical decision-making and, in the future, allow robotic military systems to make autonomous ethical judgments. We use this particular project as a case study for efforts that seek to frame morality in quantitative terms. We report preliminary results from this research, describing the assumptions and limitations of a program that assesses the relative evil of two courses of action. We compare this program to other attempts to simulate ethical decision-making, assess possibilities for overcoming the trade-off between input simplification and output reliability, and discuss the responsibilities of users and designers in implementing such programs. We conclude by discussing the implications that this project highlights for the successes and challenges of developing automated mechanisms for ethical decision making.

Keywords
decision support, ethical judgment, evil, military, modeling and simulation, robotic systems

Please address correspondence to:
Nicholaos Jones
Department of Philosophy
Morton Hall 332
University of Alabama in Huntsville
Huntsville, AL 35899 USA
email: nick.jones@uah.edu

Toward Modeling and Automating Ethical Decision Making: Design, Implementation, Limitations, and Responsibilities

1 Introductory Remarks

In November 2006, a mental health advisory team (MHAT) working for the Office of the Surgeon General published a report assessing the mental health of combat soldiers deployed in Operation Iraqi Freedom from 28 August 2006 to 3 October 2006. Part of their report addressed the topic of battlefield ethics (MHAT 2006: 34-42). Some of their findings include:

- Over 85% of Soldiers and Marines reported receiving training in how they would treat non-combatants, yet 33% of Marines and 29% of Soldiers did not agree that their commanding officers made it clear not to mistreat non-combatants (2006: 37).

- Only 47% of Soldiers and only 38% of Marines agreed that non-combatants should be treated with dignity and respect, while 17% from each group agreed that all non-combatants should be treated as insurgents (2006: 35).

- 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 (2006: 35).

- 28% of Soldiers and 31% of Marines reported facing ethical situations in which they did not know how to respond (2006: 37).

Since MHAT developed its survey questions from scratch, by virtue of never before having been tasked with an ethics assessment and failing to find relevant questions in their search of the scientific literature, there is no basis with which to compare their findings. But even without being able to compare these results to other surveys, MHAT's findings suggest that there is room for ethical improvement among Soldiers and Marines during wartime situations. MHAT recommends that soldiers be given battlefield ethics training and that behavioral health professionals serving the soldiers incorporate battlefield ethics into their counseling activities, especially when the soldiers are deployed in a combat theatre (2006: 42).

Commenting on MHAT's report, Arkin (2007) argues that autonomous robots, at some point, will be able to perform more ethically than human soldiers during combat situations. His reasons include:

- that robotic sensors for battlefield observations can surpass the capacities of human observation, and robotic processors can interpret those observations more quickly than humans;

- that robots, unlike humans, can be designed without emotions and anger that produce poor judgment;

- that robots, unlike humans, can be designed to avoid confirmation bias, distorting new information so that it fits only pre-existing beliefs (2007: 6-7).

While Arkin does not maintain that autonomous robots could be ethically perfect, he does infer that they could be ethically superior to human soldiers in combat situations (2007: 7).

One alternative to replacing or supplementing human soldiers with autonomous robotic warriors in order to improve ethical outcomes during combat situations is to remove human soldiers from combat situations by using remote-monitored telerobots such as drones (see Rozoff 2010). These robots are not capable of acting autonomously, requiring human operators to command their actions. The MHAT report notes that handling dead bodies and human remains increases the mistreatment of non-combatants by soldiers (2006: 39-41), and that and that soldiers with high levels of anger are twice as likely as soldiers with low levels of anger to engage in unethical behavior in combat situations (2006: 38). Furthermore, Mitchell (1969) found that a person's judgments regarding the hostility of others affect their rating of the comparative suitability of actions as means to ends.

One might infer from this that removing soldiers from situations that cause anger or exposure to casualties of war should decrease unethical behaviors. Sullins (2010), however, argues that removing human soldiers from combat situations apparently tends to increase unethical behaviors, confounding ethical decision making by making it more difficult for soldiers to access morally relevant situational information. This is problematic only insofar as robots do not act autonomously. For example, a recent test with drones demonstrates the possibility of autonomous robotics, with autonomous aircraft coordinating information to identify a ground target. The test

> laid the groundwork for scientific advances that would allow drones to search for a human target and then make an identification based on facial-recognition or other software. Once a match was made, a drone could launch a missile to kill the target (Finn 2011).

Since developing autonomously operating robotic systems is one of the priorities of the U.S. military (Sharkey 2008b: 87), the primary ethical challenge for autonomous robotics is designing software that allows the robots to make ethical decisions.

A second alternative to replacing or supplementing human soldiers with autonomous robotic warriors in order to improve ethical outcomes during combat situations is to provide human soldiers with support tools for ethical decision-making. This support would come in the form of software that presents human soldiers with morally salient information, taking advantage of the capacities that Arkin claims would make robotic soldiers superior to human ones: advanced perceptual and computational powers, reasoning unclouded by emotion, and judgment that is not biased by pre-existing beliefs. Unlike drones, this alternative has yet to be field-tested. Moreover, even if providing support tools to human soldiers does not increase levels of ethical behavior during combat, developing software capable of processing ethically relevant information is a prerequisite for designing fully autonomous robotic soldiers.

One of us [GR] has been involved in developing a tool for computing the "relative evil" of pairs of military actions for use by military commanders. The tool, a computer model which we shall refer to as the Metric of Evil, is designed to provide commanders with a tangible ethical viewpoint when analyzing potential courses of action by simulating the ethical judgments of human experts. This kind of simulation offers a way for commanders to explore the ethical implications of potential actions, allowing them to ask general questions (such as “What if we could reduce the number of cultural facilities destroyed?”), re-execute the tool’s programming with different inputs, and continue interacting iteratively with the tool in order to seek the most ethically viable solutions. For this reason, the tool has the potential to lead to fewer casualties, validate that a commander's decision took morality into account, and, perhaps, produce more effective military actions.

The Metric of Evil is not designed to assist soldiers with real-time ethical decision-making; nor is it designed to direct decisions of autonomous robotic systems. But it is a necessary step in those directions, and it highlights some of the significant limitations that support tools for ethical decision-making must address. For this reason, the Metric of Evil strikes us as a relevant and interesting case study about how to design robotic systems that have the capacity to reason about morality. Moreover, a longer-range hope for the metric, based upon its quantitative nature, is that it be integrated with other course of action analysis tools, thereby contributing to commanders a holistic picture of the constraints on their decisions. For, as we shall discuss, the metric is essentially a set of equations; as such, it has the potential to assist commanders in discerning an “optimal” ethical decision through sensitivity analysis, Monte Carlo simulation, a genetic algorithm, or similar methods. Our discussion of the Metric of Evil provides reasons to be cautious with respect to developments in these further directions.

We begin, in the next section, by discussing the development of the Metric of Evil, making explicit the key assumptions involved in its construction. Next, we report some results from an initial implementation of the metric, noting the significance of these results for the prospects of designing systems that have a capacity to produce ethical judgments. After highlighting some limitations of the Metric of Evil as a support tool for ethical decision making, we compare the metric to similar programs developed by McLaren (2005) as well as Anderson and Anderson (2009)**.** We orient this discussion around McLaren's thesis that any tool capable of providing ethical judgments must oversimplify its inputs in order to make ethical principles computationally tractable (2006). We then provide some remarks concerning the responsibilities associated with designing and using a tool that automates ethical decision-making and a brief report from on-the-ground experience about one of the primary challenges in discharging those responsibilities. We conclude with an overview of lessons to be learned from the Metric of Evil regarding designing and implementing automated tools for ethical decision-making.

2 Designing the Metric of Evil

In light of Tackett's (2009) proposal for a methodology to evaluate the relative amounts of "evil" associated with pairs of military events, the U.S. Army Aviation and Missile Research, Development, and Engineering Center's System Simulation and Development Directorate (AMRDEC, SSDD) tasked the Center for Modeling, Simulation, and Analysis (CMSA) and the Center for the Management of Science and Technology (CMOST) at the University of Alabama in Huntsville to refine Tackett's methodology into a useful metric and to calibrate that metric to expert evaluations of historical military events (CMSA/CMOST 2010: 62). The primary purpose of this metric is to provide a relative ethical assessments of pairs of potential military courses of action that military commanders can use as one factor in their overall course of action analyses; the secondary purposes are to reduce the amount of manpower required to provide ethical assessments for courses of action and to make explicit the implicit and unconscious priorities that produce those assessments (CMSA/CMOST 2010: 13, 15).

A conceptual prerequisite for making pairwise comparisons of the amount of evil associated with courses of actions is a working analysis of the notion of evil. Because Tackett's definition of evil as manifested intentional harm causing injury, damage or loss fails to include harms that are foreseen but not intended, CMSA/CMOST adopt a definition according to which the evil associated with an action is the intentional or anticipatable harm the action produces, where this harm includes not only harm to individual people but also damage to a society's infrastructure and violations of laws and treaties (CMSA/CMOST 2010: 9, 16). This definition, while broader than Tackett’s, does not include harm to animals and the environment. But, rather than attempt to develop a fully adequate analysis of a vague notion, CMSA/CMOST's metric design does not depend entirely upon the details of what evil is (CMSA/CMOST 2010: 48). Their product, which we shall refer to as the Metric of Evil, is a mathematical model that takes as input numerical values for observable factors relevant to the amount of "evil" associated with various military actions, and yields as output a judgment about which, if either, of two alternative courses of military action is the "lesser of two evils" (CMSA/CMOST 2010: 18).

While the Metric of Evil is designed to provide results that resemble human reasoning about morality and evil, it is not explicitly designed to do so in a way that actually resembles human reasoning. The metric simulates human ethical reasoning, because it receives as input information about factors relevant to the morality of actions and yields as output ethical judgments about those actions. However, the equations that the current version of the metric uses to convert its input to an appropriate output are not intended to represent ethical principles or logical connections between inputs in mathematical form. This distinguishes it from models like Anderson and Anderson's MedEthEx, which presumes that ethical principles can "be made precise enough to be programmed into a machine" (Anderson and Anderson 2009: 17). Some information about the Metric of Evil should help to clarify these points. (This information is taken from CMSA/CMOST 2010.)

CMSA/CMOST assume that, for each action, there is a set of potential consequences of the action relevant to the amount of evil associated with that action. They assume that these consequences are quantitative and measurable, so that for each consequence *i* there is a measurement that provides a numerical value n_i for that consequence. Estimates for high and low values for each consequence, l_i and h_i, are one set of user inputs for the Metric of Evil. A second set of user inputs are numerical values for the confidence level, c_i, associated with the chance that the actual value for the consequence *i* is somewhere within the range of its high and low estimated values. The third set of user inputs are judgments about whether the consequence is intended, anticipated, or unintended and unanticipated. (The latter is relevant only when assessing actions in hindsight.) CMSA/CMOST assume that, for each category, there is an associated numerical value m_i (measure of intentionality); these numbers are not adjustable by users and are assumed to be the same for all actions. A fourth and final set of user inputs are high and low confidence standard scores, Z_l and Z_h, representing the user's overall confidence levels regarding input values; the values for these scores are, in effect, measures of risk aversion that capture how much certainty about the metric's final output matters.

CMSA/CMOST address variances in different baseline systems of morality with three further numerical parameters. The values of these parameters can be changed to reflect different ethical priorities; but they are designed to be immune to user alteration. The first such parameter is a (normalized) weight w_i associated with each factor, such that this weight represents the importance of the factor relative to other ethically relevant factors. CMSA/CMOST assume that these weights are context-insensitive. The second parameter is the Evil Power Factor, *F*, which represents how much the intentionality of a potential consequence for an action matters to the amount of evil associated with that consequence. For example, if the number of cultural buildings destroyed is an ethically relevant consequence, a small value for *F* means that intending to destroy the building is more evil than merely foreseeing the building's destruction, while a high value for *F* means that intending to destroy the building and merely foreseeing the building's destruction produce similar amounts of evil. The third parameter is the Diminishment Factor, *D*, which represents how much each additional harm within each category of ethical relevant consequences matters to the amount of evil associated with that consequence. For example, if the number of people killed as the result of an action is an ethically relevant consequence, a small value for *D* means that killing a few people is just as evil as killing many people.

The Metric of Evil is implemented as a set of equations that takes as input numerical values for the local parameters h_i, l_i, c_i, and m_i for two courses of action *j* and *k*, as well as numerical values for global parameters w_i, Z_l, Z_h, *F*, and *D* common to both actions; calculates intermediate values for each action's mean potential evil, μ_m, and standard deviation of potential evil, σ_m (a function of c_i, Z_l, and Z_h); and yields as output the Delta Goodness, ΔG_{jk}, for the two actions. The Delta Goodness for a pair of actions is "a measure of how much less evil one [course of action] is than another" (CMSA/CMOST 2010: 28). The ethically interesting mathematics in the Metric of Evil is the equation for calculating the potential evil for an action. There are two such equations, one that provides a high estimate and one that provides a low estimate; these estimates are merged, as a function of the global parameters Z_l and Z_h, into a single assessment.

Generically, where n_i is the value associated with some consequence i, the amount of evil, E, for an action is calculated with the equation:

$$E = \Sigma\, n_i^D m_i^F w_i,$$

where the sum ranges over each potential consequence of the action. As a weighted sum, this equation is similar in structure to other decision framing models (Goodwin and Wright 2004: 43) and classical consequentialist evaluation schemes (Gips 1995: 245). The equation allows for a diminishing margin for the evil associated with increasing consequences by exponentiating the quantity n_i by the Diminishment Factor. Similarly, it allows flexibility in the significance of intention m_i by exponentiating the Evil Power Factor. The role of the equation is not to reflect how people cognitively process ethical judgments; instead, its role is to properly frame those judgments and the factors that create them.

Like McLaren's SIROCCO programs (McLaren 2003), the Metric of Evil presupposes that the ethically relevant features of different courses of action can be described in a rigorous way. While SIROCCO takes as input coded descriptions of ethical scenarios in a rigorous transcription language, the Metric of Evil takes as input numerical values about measurable potential consequences of a course of action. While SIROCCO generates output that describes considerations *relevant* to ethical assessments, the Metric of Evil generates output that *is* an ethical assessment. The equations that drive this output, however, do not purport to represent any kind of ethical principle that occurs when humans engage in ethical reasoning. Rather, the equations permit simulating the outputs of that reasoning through calibration of the values for w_i, F, and D.

If these values are adjusted properly, the Metric of Evil can output ethical judgments that match human judgments despite arriving at those judgments in a way that does not match the way in which humans arrive at their judgments. In several reviews and studies, Dawes has discussed the power of similar decision aids (1971, 1974, 1979, 1989). Her studies suggest that even models with randomly chosen weights can outperform human judges, so long as their input parameters are chosen appropriately. The primary reason for this is that people are not adept at integrating information from diverse and incompatible sources—for example, in combining students' grade point average and Graduate Record Examination scores in a meaningful way (1971) or combining more ethically-charged concerns for purposes of psychiatric diagnosis (1989). While CMSA/CMOST does not intend for the Metric to replace human decision makers (for reasons to be noted in due course), Dawes' research highlights the potential power that even simple models have to augment the decision-making process.

3 Implementing the Metric of Evil

Producing comparative judgments about the relative amount of evil associated with pairs of action using the Metric of Evil requires identifying potential consequences of actions that are relevant to the evil associated with those actions. CMSA/CMOST proposed twenty-seven such consequences: number of persons killed, wounded or injured, and captured or missing who are non-combatants, "friendly," and "enemy;"

number of non-combatants who are left without facilities that provide necessary resources to a population, who are left as homeless or refugee, who are left unemployed, and who are left economically damaged; number of infrastructure elements destroyed that are necessary for the population, that impact national or group culture, and that are otherwise present, for each of the categories non-combatant, "friend," and "enemy;" and, finally, number of minor or major violations of laws of treaties and number of national promises broken (see Table 1). Each of these consequences is measurable and relatively objective, and while some are more difficult to measure or estimate than others, the confidence level associated with each number provides a way to take into account uncertainties.

Category	Name	Unit	Unskewed Weights	Skewed Weights
Friendly force casualties	Killed	Persons	2.0%	0.0%
	Wounded or injured	Persons	1.5%	0.0%
	Captured or missing	Persons	1.0%	0.0%
Enemy force casualties	Killed	Persons	2.0%	0.0%
	Wounded or injured	Persons	1.5%	0.0%
	Captured or missing	Persons	1.0%	3.0%
Non-combatant casualties	Killed	Persons	30.0%	90.0%
	Wounded or injured	Persons	8.0%	0.0%
	Captured or missing	Persons	2.0%	7.0%
Non-combatant hardship	Left without essential facilities/resources	Persons	8.0%	0.0%
	Homeless or refugee	Persons	2.0%	0.0%
	Unemployed	Persons	1.0%	0.0%
	Economically damaged	Persons	1.0%	0.0%
Friendly infrastructure damage	Essential facilities destroyed	Count	4.0%	0.0%
	Cultural facilities destroyed	Count	2.0%	0.0%
	Non-essential facilities destroyed	Count	1.0%	0.0%
Enemy infrastructure damage	Essential facilities destroyed	Count	4.0%	0.0%
	Cultural facilities destroyed	Count	2.0%	0.0%
	Non-essential facilities destroyed	Count	1.0%	0.0%
Neutral infrastructure damage	Essential facilities destroyed	Count	8.0%	0.0%
	Cultural facilities destroyed	Count	4.0%	0.0%
	Non-essential facilities destroyed	Count	1.0%	0.0%
Moral/Ethical/Legal Considerations	Major international law violations	Count	4.0%	0.0%
	Major treaty violations	Count	2.0%	0.0%
	Minor international law violations	Count	1.0%	0.0%
	Minor treaty violations	Count	1.0%	0.0%
	National promises broken	Count	4.0%	0.0%
Global factors	Evil Power Factor (*F*)		3.00	3.00
	Low confidence range coverage factor (Z_l)		0.50	0.50
	High confidence range coverage factor (Z_h)		3.00	3.00
	Diminishment Factor (*D*)		0.85	0.5

Table 1: Optimal Weights and Parameters for Metric of Evil (CMSA/CMOST 2010: 44).

Implementing the Metric of Evil also requires assigning values to parameters meant to capture elements of a baseline morality system (weighting for each consequence, Evil Power Factor, and Diminishment Factor). CMSA/CMOST determined these values using a three-step process: first, assigning hypothetical values to each parameter; second, obtaining judgments from human experts about the relative evil associated with various pairs of historical military events; third, calibrating the hypothetical parameter values in order to maximize a match with the judgment of human experts in the test cases.

To calibrate the parameter values, CMSA/CMOST solicited ethical judgments from 35 experts, 20 of whom were Army officers, non-commissioned officers, and Army civilians, and 15 of whom were non-military religious professionals or professors with doctoral degrees in psychology, philosophy, history, and political science (2010: 72). Each expert received a packet containing detailed information, statistics, and questions on two historical case studies. (Consult CMSA/CMOST 2010 for the details of these cases.) For each case study, experts rated the relative evil of actions performed by the different groups involved in the case. For example, in the Operation Enduring Freedom case, experts judged whether the United States' actions were much more evil, more evil, neutral, less evil, or much less evil than the Taliban's actions (2010: 89). Experts then rated, on the same scale, the relative evil of actions performed by different groups involved in different cases. For example, for one packet, experts judged the relative evil of the Cuban communists' actions in the Bay of Pigs case and the United States' actions in the Operation Enduring Freedom case (2010: 90).

CMSA/CMOST mapped these ratings to the set {-2, -1, 0, 1, 2}, with -2 representing that the former action was much more evil than the latter, -1 representing that the former action was less evil than the latter, and so on. Following a modified Delphi procedure, they also assigned initial weights to the adjustable parameters in the Metric of Evil (2010: 22). After executing the Metric to obtain outputs for the metric's judgments of the relative evil of various actions, CMSA/CMOST mapped these ratings to the same set {-2,-1,0,1,2}. They then calculated an agreement rating score for the metric with a two-step procedure: first, comparing the numerical ratings of each rater to the corresponding ratings of each human expert, judging that there is agreement if the ratings had the same sign and incrementing the "agreement count" for the metric by 1 when there was agreement; and second, dividing the overall agreement count for the metric by the total number of comparisons in order to produce an agreement rating score for the metric (2010: 38-39 and 72-78). Finally, CMSA/CMOST incrementally adjusted the initial parameter weightings, calculating an agreement rating score for the metric with each adjustment until no further adjustments increased the score (2010: 42). The parameter weightings that produced the optimal agreement rating score for the metric became the "fixed" (that is, not intended to be adjusted by end users) values for the Metric of Evil. (In practice, the list of specific consequences can be modified and the parameter weightings recalibrated to represent different baseline morality systems—or to improve representation of existing ones—so long as two courses of actions are not compared with different parameter values.)

This procedure treats the collective judgments of human experts as standards of accuracy, because the Metric of Evil is intended to replicate human assessments, and because there does not appear to be any more viable standard of comparison. As CMSA/CMOST note, "there is no physics equation, authoritative data source, previously existing model, or accepted body of precedent for such assessments" that can be relied upon with certainty (CMSA/CMOST 2010: 30). In addition, a collection of human experts likely provides a broader and more useful perspective than a single expert alone.

Finally, because the numerical value associated with the mean potential evil of different actions is likely to be artificially precise, implementing the Metric of Evil requires establishing a range of values such that, for any two actions, if the Delta Goodness of those actions falls within that range, the Metric of Evil judges the actions to be *equally* evil despite having associated with them different mean amounts of potential evil (CMSA/CMOST 2010: 39).

After completing these prerequisites for implementation, CMSA/CMOST found several interesting results. We report only a few, in order to illustrate the potential for designing automated tools for ethical reasoning and to mention some of the insights gained in attempting to develop an automated tool for ethical decision-making. One result is that when all input parameters remain intact, so that no weight associated with a potential consequence can become 0% ("unskewed" weights in Table 1), the calibrated Metric of Evil produces results that agree less well with expert judgments than the results from a calibrated Metric of Evil that allows some input parameters to drop out of consideration ("skewed" weights in Table 1). This suggests that treating as relevant some of the potential consequences thought to be ethically relevant skews ethical judgments. A second result is that, when some input parameters are allowed to drop out, the output from the calibrated Metric of Evil compares favorably to the average judgments of human experts and significantly outperforms randomly generated judgments. This suggests that, while further research is still required, the methodology underlying the Metric of Evil is viable and practical (CMSA/CMOST 2010: 47). Together, these first two results support the possibility of developing a robust tool that produces ethical judgments using measurable and objective inputs, provided that it is possible to discover which factors influence the ethical judgments of human experts. This is an open topic for future research, part of which is underway and discussed below.

A third result from implementing the Metric of Evil is that the presence of non-combatant deaths is overwhelmingly a deciding factor in determining which of two actions is more evil. The calibrated metric assigns the weight of this value at 90%, which means that the number of non-combatants killed accounts for 90% of an action’s evil and thus dominates every other potential consequence of the action (see Table 1). This is a significant result, in light of MHAT's survey findings, because it highlights a discrepancy between expert and military valuations of non-combatant harm. A fourth result is that the optimal value for the Diminishment Factor, for both "unskewed" and "skewed" weightings, is less than 1.0.

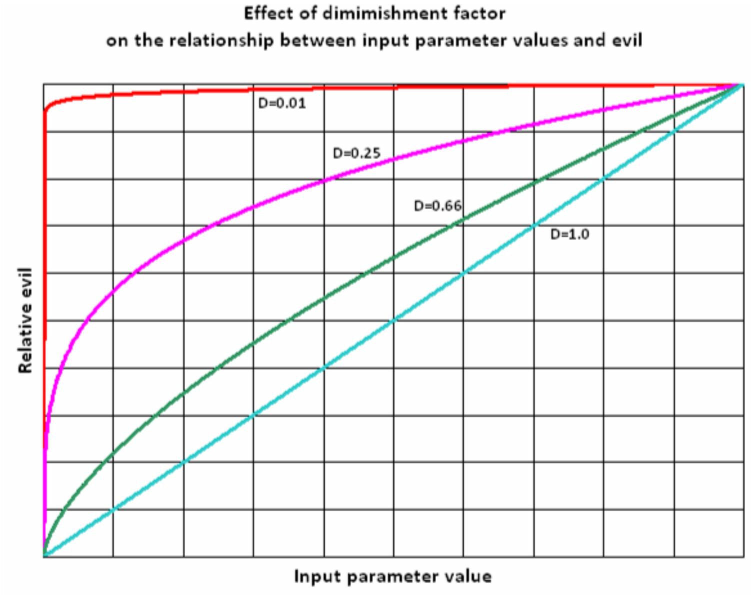

Figure 1: Effect of Diminishment Factor on Relation between Consequence and Evil (CMSA/CMOST 2010: 25)

When the Diminishment Factor is equal to 1.0, each incremental increase in the value for some potential harmful consequence produces an equally incremental increase in the amount of evil associated with that consequence; when the Diminishment Factor is less than 1.0, each increase in the value for some potential harmful consequence produces a progressively larger increase in the amount of evil associated with that consequence (see Figure 1). This suggests that the mere fact that a harm occurs is much more important than the amount of that harm that occurs. Together, these latter two results suggest that the presence or absence of non-combatant deaths is the dominating factor in many pairwise comparisons of the relative evil of military courses of action.

Whether this conclusion holds generally, or whether it is an artifact of the cases and sample size of experts used for calibration, remains an open topic for future research. The current calibration of the Metric of Evil relied upon only four case studies and judgments from 35 experts, and these limited numbers do not support strong generalizations. The preceding conclusion does, however, motivate further experiments, either with larger sample sizes or with different cases. (Such cases might include, for example, ones in which there are very few non-combatant deaths but much more significant amounts of other kinds of harm, or cases that are similar in terms of non-combatant deaths but different with respect to other factors.) Further research by [GR] is also underway, which uses the weighted sum structure of previous versions of the Metric and explores ways to balance tangible consequences and abstract principles as potential inputs.

Moreover, even without further research, the conclusion strongly supports Sharkey's recommendation that we should "severely restrict or ban the deployment of these new weapons until there have been international discussions about how they might pass an 'innocents discrimination test'" (2008b: 89). It also makes pressing his concern that "no computational system can discriminate between combatants and innocents in a close-encounter contact" (2008a: 1801). Morrow's assessment of the potential for disconnect between technology and morality, that "[t]he story of evil in the world is so often a matter of hardware outperforming conscience: Can outruns Should. Or rather, Can outruns Should Not" (2003: 56), corroborates this concern.

4 Limitations of the Metric of Evil

While the results from implementing the Metric of Evil provisionally support the viability of supporting or automating ethical decision-making with computer simulations, and while they further support a thesis about the relative importance of non-combatant harm to the morality of military actions, there remain significant limitations to using the Metric of Evil as a support tool for ethical decision-making. These limitations include but are not limited to: the comparative nature of the metric's judgments; the moral relativity of those judgments; the meaning of the judgments; the reliability of the judgments; the authority of the judgments; and the objectivity of the judgments.

First, the Metric of Evil only ranks pairs of actions as more or less evil than each other. Because one action can be less evil than another and yet still be morally impermissible, the metric does not yield as output absolute moral judgments concerning whether a particular course of action is right or wrong. Nor is it intended to do so. The context for which the metric is designed (supporting military command decisions) is one in which the ultimate goal is to perform one of the courses of action under consideration. The Metric of Evil produces a judgment about the relative evil of two potential courses of action, and the transitivity of the *is less evil than* relation supports inferences from sets of such pairwise comparisons to rankings of the relative evil of arbitrarily many potential courses of action. (If action A is less evil than action B and action B is less evil than action C, then action A is less evil than action C.) However, there is no direct way to infer absolute moral judgments from these comparative judgments.

The primary cause of this limitation is that neither the metric's inputs nor its constituent equations represent any sort of absolute moral code or principle. The inputs include factors thought to be ethically relevant; and the central equations support output that is calibrated to conform to expert human judgments but not explicitly based upon codes or principles that support those judgments. Accordingly, developing the Metric of Evil into a tool that supports ethical decision-making in a more robust sense requires either finding a way to mathematize moral codes or, perhaps, finding paradigm cases of morally permissible and impermissible actions to which potential courses of action can be compared. Such cases would support inferences from comparative moral judgments to absolute moral judgments in much the same way that, were the world Newtonian, information about reference frames that are stationary relative to absolute space would support inferences about the absolute state of motion of objects. For any action less evil than a paradigm case of a morally permissible action is morally permissible, and any action at least as evil as a paradigm case of a morally impermissible action is morally impermissible.

The challenge, of course, is to identify the paradigm cases. This requires identifying actions that experts tend to judge similarly, and which also permit meaningful comparisons to the courses of action of interest to users of the Metric of Evil. Because the *Belmont Report* (1979) identifies paradigm cases of this sort from the realm of bioethics, it seems that there is no *apriori* reason to suppose that such cases do not exist in the realm of military ethics. If such cases can be found, the absolute moral judgments they help to support would increase the Metric of Evil's range of application, helping military commanders to decide not merely which course of action is the lesser of available evils but also whether to pursue any considered course of action or, instead, search for more options.

A second limitation of the Metric of Evil is that the judgments it produces are calibrated toward a particular baseline morality. Again, this is intentional. Because the metric is intended to support American military operations, it is designed to be "a model of the ethical values of the analyst using the tool or of the society considering the courses of action being evaluated" (CMSA/CMOST 2010: 30). Accordingly, while the displayed output of the metric takes the form 'Action A has the same/more/less evil than action B,' there is an implicit qualification to this output, namely, 'relative to a particular system of morality.' As noted, the design of the metric permits flexibility in choosing appropriate consequences (or other inputs) and calibrating the metric to the baseline morality of different groups of experts. But further research would be required to determine whether the metric can be validated with respect to experts who rely upon other moral systems. Moreover, if the goal of automated ethical decision-making is to produce a robot capable of making non-relativized judgments about morality (whether absolute or comparative), further research is required to determine whether there is some sort of universal baseline morality relative to which the Metric of Evil can be calibrated.

A third limitation of the Metric of Evil concerns the meaning of its (comparative and relativized) ethical judgments. The concept of evil is notoriously difficult to define in a crisp, clear manner. Reasoning that evil is the "most powerful word in the language, and the most elusive" and that it is "felt rather than understood," Morrow claims that evil is not "subject to measurement and scientific inspection" (2003: 7, 28, 35); that, since "evil is evil," evil acts cannot be compared (2003: 83-84). Despite this, attempts have been made. For example, Philip Zimbardo describes Stanley Milgram's famous experiment in obedience as an attempt to quantify evil by tying it to the sheer number of electrical shocks a subject is willing to administer to an actor under direct orders (Zimbardo 2004: 27). He notes that, in contrast to bad apples that may spoil the whole barrel, evil is like vinegar that will "always transform sweet cucumbers into sour pickles," adding rhetorically that, for this reason, the vinegar ought to be understood (Zimbardo 2004: 47). Moreover, Baron-Cohen (2003) claims that a scientific understanding of evil is required in order to avoid circular reasoning and other forms of fallacious reasoning about the subject. Welner's Depravity Scale aims to provide accountability and clarity in using definitions such as "evil," "depraved" or "heinous" in the courtroom, primarily because judging crimes as such affects sentencing (Welner 2007).

Studying the concept of evil in a quantitative and methodical fashion is one way to address the challenge of crisply defining and framing evil, especially if a concern of military commanders is minimizing the amount of evil their decisions produce (or optimizing the ethical impact of their actions). Undoubtedly, there is contention and disagreement over how evil ought to be defined, and CMSA/CMOST intend for the Metric to capture the notion of evil as closely as possible. However, even if the Metric captures only a related or tangential notion, such as intentional harm, it still might benefit commanders if the boundaries of the Metric and the limitations of its actual domain of measurement are clearly understood, for the primary goal of the tool, regardless of what exactly is being measured, is to help military commanders make ethical decisions.

CMSA/CMOST chose to use a definition of evil involving the notion of intentional or anticipatable harm. While this definition is relatively suitable in the context of military operations, it is not clear that it is entirely adequate. Significantly, some human raters polled in CMSA/CMOST's efforts to calibrate the Metric of Evil, especially those with a background in psychology or ethics, expressed difficulty in working with the definition of evil as intentional or anticipatable harm; some also substituted their own definition of evil (CMSA/CMOST 2010: 46). These responses have the potential to compromise the results of calibration. If the metric is calibrated to human judgments based upon competing definitions of evil, it is not clear what the outputs from a calibrated metric mean. (A better definition of evil might alleviate these particular issues; then again, it might lead to other problems, such as undesired variability in the metric's results. Moreover, if there are problems with all definitions of evil, it might not be possible to calibrate a computer model like the Metric of Evil in a meaningful way.)

A fourth limitation of the Metric of Evil concerns the reliability of its output. Unlike models of decision-making in game theory, the Metric of Evil does not rely upon assumptions that actors are rational and self-interested (see Dixit 2004). However, like those models, the Metric of Evil explicitly ignores factors that are difficult to quantify. Intuition and bias are especially difficult to represent in a mathematical fashion; but Rogerson et al. argue that ignoring these factors is a dramatic limitation of models designed to aid ethical decision-making (2011: 614). Factors that affect an action's overall context, such as actors' motivations and political situations, are also difficult to define and quantify. If these additional factors are significant to ethical judgments within particular moral systems concerning pairwise comparisons of evil, a Metric of Evil that is calibrated to be reliable in some contexts is not necessarily reliable in all contexts. This is a topic for future research. It might be that those who use the tool need to weigh context themselves.

A fifth limitation of the Metric of Evil concerns the authority of its output. One of the main reasons that tools like the Metric of Evil fail is that their users do not trust the outputs. As Bell notes,

> A system which simply says “do this” or “the answer is 42” without any justification or explanation is unlikely to find acceptance. If the user can see why the system produced the results it did, then acceptance is much easier (1985: 617).

The Metric of Evil produces, as output, *only* a comparative judgment about the relative evil of various potential courses of action. But, unlike other ethical support tools such as McLaren's Truth-Teller and SIROCCO programs, it provides no explanation for this output. Moreover, because the Metric of Evil's underlying equations do not represent elements involved in human ethical reasoning, the metric is not equipped to allow users to extract explanations that help them to understand why an expert might provide similar judgments: even if users calculate the output for themselves rather than allow the software perform the calculation, their calculations rely on equations and parameter values that have no human significance. Accordingly, even if the output of the Metric of Evil is reliable, in practice that output is likely to lack authority as something upon which humans should base their decisions. Of course, some humans (the technologically credulous ones) are inclined to believe anything a computer says; and here the danger is not that such users will fail to trust the Metric of Evil's output, but that they will trust that output uncritically (Bell 1985: 618). Because the Metric of Evil does not explain its outputs, it does not provide these more credulous users with materials to check for exceptions to whatever general reliability the Metric of Evil's outputs happens to have. Given the intended short-term applications of the Metric of Evil, and especially the potential that military commanders might rely upon the metric to make decisions that affect human lives, this limitation strikes us as having special moral significance.

One final limitation of the Metric of Evil concerns the objectivity of its judgments. Despite the metric's attempt at objectively framing military ethical concerns, some subjectivity remains in the categorization of inputs. For example, Tackett's preliminary version of the metric requires analysts to provide estimates for harms along a Harm Index which indicates increases in harm severity. The preliminary version also requires analysts to provide an estimate of Hardships and Sufferings. Analysts must further rate each consequence of an action according to an "Order of Evil"—analysts inform the model whether each consequence is Necessary, Consequential, Selfish, or Malicious (see Tackett 2009). While each of these categories along the Harm Index and the Order of Evil are defined, there is some inherent subjectivity present in categorizing harms based upon these definitions. This subjectivity provides a great amount of flexibility. This is useful, insofar as it allows analysts to incorporate context to a certain extent (an action committed in self-defense might be seen as less evil than the same action committed with no such context or justifiability). However, an analyst could, consciously or not, project personal biases into categorizations that deem certain consequences to be necessary and others to be malicious, essentially justifying biased decisions by pointing to the results of the tool's analysis. Hence, if not properly framed by the tool, subjectivity can lead to inconsistency in results or, worse, allow the tool to be used improperly in a way that justifies unreasonable or immoral actions.

The CMSA/CMOST-developed version of the metric attempts to frame harms more objectively than Tackett's preliminary version. Factors previously lumped together as hardships or sufferings are more explicitly enumerated into categories such as destroying essential facilities, leaving civilians unemployed, and breaking international treaties. The factors used by the model do not represent a comprehensive list of potential ethically-charged consequences for actions—and, of course, in reality, such a list is virtually infinite—but CMSA/CMOST view the factors chosen as among the most relevant. In addition, rather than using the Order of Evil categories present in the preliminary version of the metric, CMSA/CMOST's version classifies consequences as intended, merely anticipatable, or completely unanticipated. There remains a certain level of subjectivity in this scheme, however—for example, the difference between “major” and “minor” treaty violations is not crisply defined. Overall, however, this newer version of the metric presents a more objective approach with clearly defined harms. While it thereby allows for more systematic, repeatable, verifiable, and potentially “honest” analyses, it also might sacrifice some of the flexibility inherent in its predecessor’s design (regarding, for instance, the ability to take into account contextual factors). This tradeoff illuminates the sheer difficulty in properly framing ethics in a way that can be systematically understood and modeled while still accounting for all of the factors necessary to form a complete picture surrounding an action to be studied. Additionally, it underscores the challenges to be anticipated by attempts to responsibly implement a capacity for ethical decision-making in autonomous robotic systems.

5 Comparisons to Truth-Teller, SIROCCO, and MedEthEx: McLaren’s Tradeoff

The Metric of Evil differs from more sophisticated tools, such as MedEthEx, primarily in terms of approach, simulating rather than mimicking the processes that humans use to produce ethical judgments. The primary reasons for this are that the Metric of Evil aims to allow military commanders to include ethical assessments of potential actions in their overall course of analyses, where many opposing factors must be taken into account, and that the metric is designed to facilitate exploration of descendant tools capable of acting as core modules that can be connected to a larger software chain to provide military commanders with a holistic perspective of various military scenarios. Given this potential development of the Metric of Evil, it is important to have some perspective on the significance of the metric’s aforementioned limitations. Accordingly, this section compares the Metric of Evil with some other support tools for ethical decision-making.

McLaren’s Truth-Teller and SIROCCO programs, as well as Anderson and Anderson’s MedethEx program, avoid the authority limitation of the Metric of Evil. McLaren’s programs take as input suitably coded descriptions of ethical scenarios and yield as output ethically relevant features of those scenarios intended to “stimulate the moral imagination” of program users in order to help them make decisions (McLaren 2006). Because these outputs do not provide ethical judgments about the moral status of the central actions in ethical scenarios, these programs leave their users as the ultimate authorities for judging the significance of the identified features. Anderson and Anderson’s MedEthEx program, unlike McLaren’s programs, yields as output judgments about whether particular actions are morally permissible; but, unlike the Metric of Evil, MedEthEx allows its users to view the ethical principles that support its output judgments (Anderson and Anderson 2009). This explanation of the output enhances the authority of the program’s judgments. For, as Bell notes, “[i]f the user can see why the system produced the results it did, then acceptance [of the results] is much easier (1985: 617).

While lack of output authority is a significant practical and moral limitation of the Metric of Evil as a support tool for ethical decision-making, and while the absence of this limitation is a definite advantage for programs such as Truth-Teller, SIROCCO, and MedEthEx, lack of output authority is not a significant disadvantage of the Metric of Evil with respect to implementing some descendant of the Metric of Evil as a program that provides ethical guidance to autonomous robotic systems. Compliance with a program's output judgments would be automatic in a robotic system, built into the robot's programming rather than mediated by human decision. Hence, the absence of an explanation or justification for the output of an "ethics" module in an autonomous system is not a barrier to translating the output into action. Accordingly, with respect to this potential future application of a descendant of the Metric of Evil, other limitations of the Metric of Evil, and especially limitations concerning output reliability, become more significant.

McLaren's programs avoid many of these other limitations, merely by virtue of the nature of their outputs. Because the programs produce only judgments of ethical relevance, the outputs are neutral with respect to whether ethical considerations taken to be important from other systems of morality are *also* relevant. This avoids limitations with respect to moral relativity and reliability. However, because the outputs do not provide ethical judgments about the moral status of actions, and because ethical decision-making requires such judgments, these programs cannot be extended in ways that would automate ethical decision-making. McLaren seems to accept this tradeoff, maintaining that any tool that provides ethical judgments must oversimplify its inputs in order to make ethical principles computationally tractable, while any tool that avoids oversimplification can provide ethically relevant information but not judgments (2006). We shall refer to this thesis as *McLaren's Tradeoff*.

The limitations of the Metric of Evil support McLaren's Tradeoff, and they help to clarify McLaren's notion of input oversimplification. The user-inputs to the metric are numbers for various factors: number of destroyed facilities, number of non-combatant deaths, number of treaty violations, and so on. These numbers are not simplified in any obvious sense. Instead, the simplifications occur in specifying the factors, categorizing the numbers for those factors, and determining the moral significance of the numbers. These simplifications are *oversimplifications* by virtue of leading to limitations concerning reliability, objectivity, and moral relativity, respectively. If McLaren's Tradeoff is correct, Anderson and Anderson's MedEthEx program should exhibit similar oversimplifications. We shall argue that it does.

MedEthEx program takes as input judgments regarding whether competing courses of medical action violate or respect different *prima-facie* duties that medical professionals owe to their patients, and it yields as output a judgment about whether a particular course of action is morally permissible. This output judgment is produced by a program that ranks the relative importance of various *prima-facie* duties, and the rankings are calibrated by machine learning techniques with reference to expert judgments about the relative importance of those duties in uncontroversial ethical scenarios (see Anderson and Anderson 2009). Accordingly, MedEthEx differs from the Metric of Evil in (at least) five respects: the output judgments are absolute rather than comparative; the program that produces those judgments is grounded in principles about *prima-facie* duties rather than stark expert assessments; this program is calibrated using machine-learning techniques rather than survey-based experiments; and the inputs to the programs are judgments about whether duties are respected or violated rather than judgments about numerical quantities. Despite these differences, MedEthEx exhibits oversimplifications similar to the Metric of Evil's. For MedEthEx exhibits similar limitations with respect to the moral relativity, reliability, and objectivity of its outputs.

The Metric of Evil exhibits limitations with respect to reliability of its outputs by virtue of relying upon a potentially incomplete list of morally relevant factors. MedEthEx is *also* calibrated to a particular baseline morality, namely, the list of *prima-facie* duties from Beauchamp and Childress's *Principles of Biomedical Ethics* (1979). While Beauchamp and Childress present their work as the consensus view of biomedical ethicists, DeMarco (2000) argues that their list of duties is incomplete. According to DeMarco, in addition to the *prima-facie* duties in *Principles of Biomedical Ethics*, there is also a *prima-facie* duty he calls the mutuality principle: "Act to establish the mutual enhancement of all basic values" (2000: 102). If these additional duties are significant to ethical judgments, their omission from MedEthEx compromises the reliability of that program's outputs. Anderson and Anderson, aware of this limitation, note that MedEthEx can be updated to incorporate further duties in its analyses (2009: 19).

While updates to MedEthEx's list of *prima-facie* duties might enhance the reliability of MedEthEx's outputs, they do not ameliorate the moral relativity of those outputs. For, even supposing a complete list of *prima-facie* duties, judgments about whether those duties are violated or respected in particular cases vary with context. For example, according to Fan, while judgments about autonomy tend to focus on individual independence and self-determination in "Western" contexts, "East Asian" contexts emphasize, instead, family-determination and harmonious dependence (1997; see also Holm 1995). The judgments that MedEthEx produces are, accordingly, relative to whatever baseline morality operative is operative in producing the data points for program calibration. For example, Anderson and Anderson note that, using machine learning techniques, MedEthEx discovered a general ethical principle:

> *A doctor should challenge a patient's decision if it isn't fully autonomous and there's either any violation of nonmaleficence [the duty not to harm unnecessarily] or a severe violation of beneficence [the duty to benefit without unnecessary harm]* (2009: 18).

Insofar as the data points used for machine learning rely upon a "Western" interpretation of patient autonomy rather than an "East Asian" one, this discovery is relative to a "Western" baseline morality.

Finally, quite apart from limitations concerning the completeness of the list of *prima-facie* duties and the interpretive variance of those duties across cultural contexts, there is a degree of subjectivity in MedEthEx's outputs. The input to MedEthEx is information about whether particular duties have been respected or violated. But obtaining this information is not a straightforward procedure. Beauchamp and Childress's duties achieve generality at the cost of abstraction, so that giving the duties traction with actual ethical scenarios requires specifying their content (see DeMarco 2000; Holm 1995). For example, determining whether an action violates a patient's autonomy or whether an action potentially leads to a preventable harm requires giving more content to notions of *autonomy*, *preventable*, *harm*, and so on. There are not formulaic ways to do this, and MedEthEx provides no guidance for how to do so. Because these determinations can vary among users (or programmers), the need to specify the content of *prima-facie* duties renders the output of MedEthEx less than fully objective.

Anderson and Anderson, aware at least of the reliability limitation of MedEthEx, counsel that "*we should probably not allow machines to engage in actions where there is not a consensus among ethicists as to the correct way to behave*" (2009: 19). Even if the requirement of consensus among ethicists, across varying and often incompatible systems of morality, is an unrealistic standard, the advice is well taken. Both MedEthEx and the Metric of Evil highlight the challenges in designing programs that yield reliable, objective, absolute, and unqualified ethical judgments. McLaren's Tradeoff casts doubt on the possibility of such a program. (Whether this doubt is significant depends, in part, upon whether the standard for success is designing a program that performs better in making moral judgments than human "programs.")

Whether MedEthEx or the Metric of Evil can be revised in ways that remove or ameliorate their current limitations, and thereby circumvent McLaren's Tradeoff, remains a question for continued research. Regarding the Metric of Evil, we are not aware of any conclusive reason to suppose that the program cannot be improved in ways that significantly ameliorate its current limitations. (We are not as familiar with MedEthEx.) Consider, for example, the objectivity of the Metric of Evil's outputs. The main impediment here is finding a way to remove the subjectivity associated with categorizing inputs to the Metric of Evil. Welner (2007) suggests a strategy for removing the kind of subjectivity. Welner's concern is to develop an objective standard for whether a criminal action is depraved. His methodology involves three stages: first, identify features that tend to be associated with actions typically classified as depraved; second, develop a standard, based upon those features, that does not require subjective judgments for its application; and third, conduct a survey to determine whether there is a consensus of support from the general public regarding the way in which that standard classifies actions. A similar methodology might be promising for developing a more objective standard for whether, say, treaty violations are "major" or "minor." (Continuing research beyond the CMSA/CMOST effort (by [GR]), still in its infancy, uses the core structure of a simple linear model but focuses on finding ways to standardize presentation of inputs in a way that circumvents McLaren's Tradeoff to a greater degree. An expected benefit of this research is an assessment of how military and non-military experts weigh certain ethical principles and consequences.)

6 Responsibilities of Designers and Users

Even if McLaren's Tradeoff is ultimately unavoidable, the Metric of Evil is not entirely useless. It outputs *something*: a comparative judgment of the relative evil of two courses of action, where the judgment is relative to a particular baseline morality, insensitive to contextual factors, and potentially biased by a degree of subjectivity from its inputs. This kind of output might prove to be useful for standalone applications or as part of a mechanism for reasoning in combat machines. For example, even if the metric's outputs are not treated as guidance for ethical decision-making, they might serve as useful prompts for military analysts to reconsider decisions prior to action when those decisions conflict with the metric's output or explore the ethical "space" surrounding their potential courses of action, and they might help to direct the attention of decision-makers to unnoticed or underappreciated ethical factors.

If the Metric of Evil's outputs prove to be useful in some way, the metric's limitations must be well-understood by users if the metric is to provide actionable results. If users misunderstand the conceptual framework behind the metric, misinterpret its output, or blindly use its output without tempering it through an application of their own ethical judgments, the metric has the potential to justify completely immoral actions. Moreover, given Morrow's poignant reservations about classifying and quantifying evil, the limitations inherent in any attempt to quantify and scientifically evaluate evil should be made absolutely clear before people use tools that such attempts produce. Responsibility falls on both the designers and the users of the metric to ensure that it is used properly. While users must not deploy the metric in an irresponsible way, the designers of the metric should make its limitations clear to the users.

While the Metric of Evil has been incorporated into prototype software, CMSA/CMOST's concern regarding their responsibilities as designers has led them to abstain from deeming the metric suitable for practical use. Nonetheless, their experience in designing more mature decision support tools have revealed issues that may arise in communicating to users the limitations of deploying a tool for automated ethical decision-making. (One of us [GR] has been involved directly in such efforts with the Charger Nursing Dashboard, a software package designed to help nurse managers make decisions for their hospital units. Underlying the package is a set of equations relating patient information, overall information about a hospital unit and staffing data to various medical and financial outcomes for the unit. For more information, see Anderson et al. 2011.)

One of the primary issues in communicating the limitations of using a decision-making support tool for is that it is difficult to convey to users the appropriate level of confidence warranted in the tool's output. Users seem to be to inclined toward innately, and naively, interpreting output that provides a "go/no-go" response or reduces a forecast to a simple "green/red" display as an absolutely certain conclusion. Ornstein foresaw this problem for military applications long ago (in terms of technological advancement), writing in 1987 that

> if the only role of a human participant is to watch a meter and push a button when the needle goes from green to red, then the participation is merely symbolic. Under such circumstances reflection and judgment are effectively eliminated, and these are precisely the qualities that constitute the crucially important human contribution (Ornstein 1987: 9-10).

As we have noted, this inclination toward "merely symbolic" interpretation would be especially dangerous in the context of military applications of decision support tools, and while users of such tools have a responsibility to avoid this kind of oversimplification, designers of the tools have a responsibility to forestall user misunderstandings by providing clear guidelines and intuitive user interfaces.

Unfortunately, at present we have no well-tested idea for how to address this kind of challenge. Determining how to fulfill these responsibilities remains an open research project.

7 Concluding Remarks

There is every indication from the U.S. government that developing autonomously operating robotic systems for military applications is a high priority for the near future (see U.S. Department of Defense 2007). Given that the U.S. Army provided funding for research on the Metric of Evil, there is good reason to suppose that those involved in military operations planning have identified a need for models that properly frame ethical concerns in military contexts. These models potentially include support tools for operation planning, to be implemented for providing decision-making guidance to military commanders. Eventually, they also might include tools for use during military engagements, to identify ethical constraints in the reasoning and decision-making processes of automated robotic systems that have identified potential targets.

The Metric of Evil is relevant as a case study for the assumptions and challenges involved in designing and implementing a support tool for ethical decision-making in military contexts. Undoubtedly, the methodology behind the Metric of Evil requires improvement before it is robust enough to produce a tool that provides responsible ethical guidance for operation planning, much less a tool that responsibly automates ethical decision making in robotic systems. We have discussed several limitations surrounding the current design and implementation of the Metric of Evil. Research to improve the metric on several fronts is underway, including an enhanced calibration experiment, an improved model design, and a more rigorous treatment of ethical issues regarding designer and user responsibilities. For the moment, however, the metric demonstrates the possibility of developing a tool that supports military commanders or autonomous robotic systems in making ethical judgments. Moreover, some preliminary results of implementing the metric highlight the importance of developing computable criteria that allow robotic systems to distinguish combatants from non-combatants. This topic should be explored rigorously and extensively before deeming the metric ready for widespread use on the battlefield.

Acknowledgements

[Omitted for blind review.]

References

Anderson, E.F., K.H. Frith, and B. Caspers. (2011). "Linking Economics and Quality: Developing an Evidence-Based Nurse Staffing Tool." *Nursing Administration Quarterly* 35: 53-60.

Anderson, S.L. and M. Anderson. (2009). “How Machines Can Advance Ethics.” *Philosophy* Now 72: 17-19.

Arkin, R.C. (2007). *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*. U.S. Army Research Office Technical Report GIT-GVU-07-11.

Baron-Cohen, S. (2003). *The Science of Evil: On Empathy and the Origins of Cruelty*. New York: Basic Books.

Beauchamp, T.L. and J.F. Childress. (1979). *Principles of Biomedical Ethics*. New York: Oxford University Press.

Bell, M.Z. (1985). "Why Expert Systems Fail." *The Journal of the Operational Research Society* 36: 613-619.

Belmont Report (1979). *Belmont Report—Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Available online at <http://ohsr.od.nih.gov/guidelines/belmont.html>.

Center for Modeling, Simulation, and Analysis, and Center for the Management of Science and Technology [CMSA/CMOST]. (2010). "Developing and Calibrating a Quantitative Metric of Evil for Use in Course of Action Analysis." *Final Technical Report*. AMREDC, SSDD.

Dawes, R.M. (1971). "A Case Study of Graduate Admissions: Applications of Three Principles of Human Decision Making," *American Psychologist* 26: 180-188.

Dawes, R.M. (1979). "The Robust Beauty of Improper Linear Models in Decision Making," *American Psychologist* 34: 571-582.

Dawes, R.M. and B. Corrigan. (1974). "Linear Models in Decision Making," *Psychological Bulletin* 81: 93-106.

Dawes, R.M., D. Faust, and P.E. Meehl. (1989). "Clinical Versus Actuarial Judgment," *Science*, 243: 1668-1674.

DeMarco, J.P. (2000). "Principalism and Moral Dilemmas: A New Principle," *Journal of Medical Ethics* 31: 101-105.

Dixit, A.K. and S. Skeath. (2004). *Games of Strategy*, second edition. New York: W.W. Norton & Company, Inc.

Fan, R. (1997). "Self-Determination vs. Family-Determination: Two Incommensurable Principles of Autonomy: A Report from East Asia," *Bioethics* 11: 309-322.

Finn, P. (2011). "A Future for Drones: Automated Killing." *Washington Post*. 19 September.

Gips, J. (1995). "Towards the Ethical Robot." In K. Ford, C. Glymour, and P. Hayes (eds.), *Android Epistemology* (Cambridge: MIT Press): 243-252.

Goodwin, P. and G. Wright. (2004). *Decision Analysis for Management Judgment*, third edition. West Sussex: John Wiley & Sons Ltd.

Holm, S. (1995). "Not Just Autonomy—The Principles of American Biomedical Ethics," *Journal of Medical Ethics* 21: 332-338.

McLaren, B.M. (2003). "Extensionally Defining Principles and Cases: An AI Model." *Artificial Intelligence Journal* 150: 145-181.

McLaren, B.M. (2005). "Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning." *Papers from the AAAI Fall Symposium*, Technical Report FS-05-06: 70-77.

McLaren, B.M. (2006). "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions." *IEEE Intelligent Systems* 6: 2-10.

Mental Health Advisory Team (MHAT) IV, Office of the Surgeon General. (2006). *Final Report: Operation Iraqi Freedom* 05-07.

Mitchell, T.O. (1969). *Observer's Hostility as a Factor in Judgments of Behavior in Hostility-Provoking Situations*. Ph.D. Dissertation.

Morrow, L. (2003). *Evil: An Investigation*. New York: Basic Books.

Ornstein, S. M. (1987). “Computers in Battle: A Human Overview.” In D. B. Bellin and G. Chapman (eds.), *Computers in Battle—Will They Work?* (Orlando: Harcourt Brace Jovanovich, Inc): 1-43.

Rogerson, M.D., M.C. Gottlieb, M.H. Handelsman, S. Knapp, and J. Younggren. (2011). "Nonrational Processes in Ethical Decision Making." *American Psychologist* 66: 614-623.

Rozoff, R. (2010). “Decade of the Drone: America's Aerial Assassins.” *Global Research*.

Sharkey, N. (2008a). “The Ethical Frontiers of Robotics.” *Science* 322: 1800-1801.

Sharkey, N. (2008b). "Grounds for Discrimination: Autonomous Robot Weapons." *Rusi Defense Systems* 11: 86-89.

Sullins, J.P. (2010). "RoboWarfare: Can Robots Be More Ethical than Humans on the Battlefield?" *Ethics and Information Technology* 12: 263-275.

Tackett, G.B. (2009). "Framework for Quantification of Evil as a Metric For Course of Action (CoA) Analysis." Draft Technical Report. AMRDEC, RDECOM.

U.S. Department of Defense. (2007). *Unmanned Systems Roadmap 2007-2032*.

Welner, M. (2007). “Classifying Crimes by Severity: From Aggravators to Depravity.” In J. Douglass, R. Ressler, and A. Burgess (eds.), *A Crime Classification Manual* (Jossey-Bass): 55-72.

Zimbardo, P.G. (2004). “A Situationist Perspective on the Psychology of Evil: Understanding How Good People are Transformed into Perpetrators.” In A.G. Miller (ed.), *The Social Psychology of Good and Evil* (New York: Guilford Press): 21-50.