Designing and Implementing a Model for Ethical Decision Making

1 Introductory Remarks

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.

Our plan is, first, to discuss the development of the Metric of Evil, making explicit the key assumptions involved in its construction. Then, second, 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.

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.

The Metric of Evil presupposes that the ethically relevant features of different courses of action can be described in a rigorous way. The Metric of Evil takes as input numerical values about measurable potential consequences of a course of action, and 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.

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.

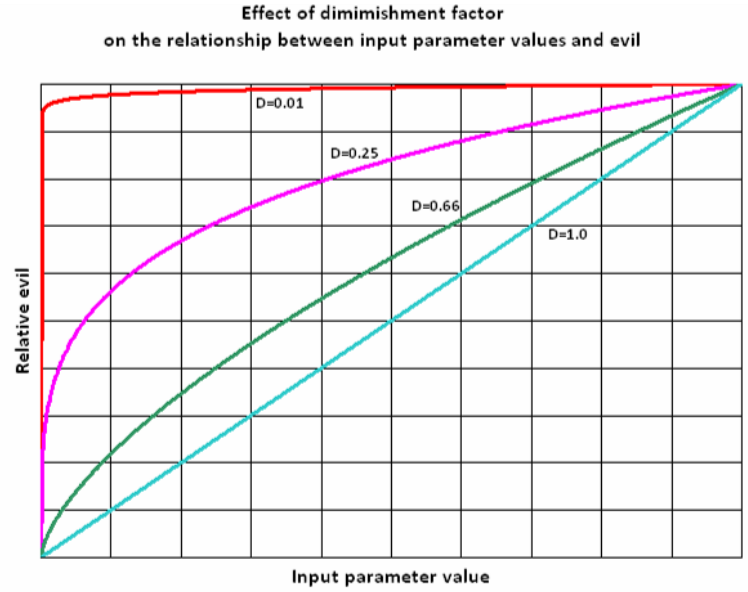

Figure 1: Effect of Diminishment Factor on Relation between Consequence and Evil (CMSA/CMOST 2010: 25)

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.

4 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. Research to improve the metric on several fronts is underway, including an enhanced calibration experiment, an improved model design, and efforts to address limitations which we have discussed in other work. 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.

References

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

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.

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.

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*.