Applications

JDMS

Journal of Defense Modeling and Simulation: Applications, Methodology, Technology
2016, Vol. 13(2) 195–211
© The Author(s) 2015
DOI: 10.1177/1548512915581213
dms.sagepub.com
SAGE

A principles-based model of ethical considerations in military decision making

Gregory S Reed[1], Mikel D Petty[1], Nicholaos J Jones[1], Anthony W Morris[2], John P Ballenger[1] and Harry S Delugach[1]

Abstract

When comparing alternative courses of action, modern military decision makers often must consider both the military effectiveness and the ethical consequences of the available alternatives. The basis, design, calibration, and performance of a principles-based computational model of ethical considerations in military decision making are reported in this article. The relative ethical violation (REV) model comparatively evaluates alternative military actions based upon the degree to which they violate contextually relevant ethical principles. It is based on a set of specific ethical principles deemed by philosophers and ethicists to be relevant to military courses of action. A survey of expert and non-expert human decision makers regarding the relative ethical violation of alternative actions for a set of specially designed calibration scenarios was conducted to collect data that was used to calibrate the REV model. Perhaps unsurprisingly, the survey showed that people, even experts, disagreed greatly amongst themselves regarding the scenarios' ethical considerations. Despite this disagreement, two significant results emerged. First, after calibration the REV model performed very well in terms of replicating the ethical assessments of human experts for the calibration scenarios. The REV model outperformed an earlier model that was based on tangible consequences rather than ethical principles, that earlier model performed comparably to human experts, the experts outperformed human non-experts, and the non-experts outperformed random selection of actions. All of these performance comparisons were measured quantitatively and confirmed with suitable statistical tests. Second, although humans tended to value some principles over others, none of the ethical principles involved—even the principle of not harming civilians—completely overshadowed all of the other principles.

Keywords

Human behavior modeling, decision analysis, machine ethics, modeling and simulation

1. Introduction and motivation

Large-scale force-on-force conflicts are becoming much less frequent. Modern militaries more often must execute operations that include asymmetric warfare, counterinsurgency, and nation-building, which may involve difficult ethical issues, e.g., an adversary's use of non-combatants as cover. Today a military commander may be faced with an operational situation in which some action must be taken and two or more actions are available, but all potential actions may have negative ethical consequences. Due to the complex nature of the situations now facing military decision-makers, a more comprehensive picture—one that encompasses the ethical implications of military actions—needs to be developed to facilitate sound decision-making.

Machine ethics, an emerging field, involves developing ''machines'' (either as tangible hardware or as mathematical or logical models) with ethics codified as principles, parameters, and procedures, allowing them to consider the ethical implications of potential actions. This research

[1]University of Alabama in Huntsville, Huntsville AL, USA
[2]Army Research Laboratory, AMCOM Field Element, Redstone Arsenal, AL, USA

Corresponding author:
Gregory S Reed, University of Alabama in Huntsville, Shelby Center 142, 301 Sparkman Drive, Huntsville, AL 35899, USA.
Email: gregory.reed@uah.edu

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

involves the development of a new mathematical model specifically applied to military courses of action (COAs). That model, referred to as the Relative Ethical Violation (REV) model, is designed to help military decision makers analyze the types of operations they now must conduct. To be clear, the REV is not modeling the entire military decision-making process. A commander chooses a COA based on several considerations, including military effectiveness, logistical feasibility, and ethical violation. The REV model is focused only on the ethical concerns that influence and constrain modern military operations. The model is intended to support military decision-making by evaluating the ethical implications of potential COAs.

Although the development and evaluation of the REV model was influenced and informed by the authors' earlier work on a similarly intentioned model known as the Metric of Evil (MOE), which is summarized in Section 2, the REV model is based on a completely different perspective on military ethics. The earlier MOE model was focused on the tangible, quantifiable *consequences* of military COAs, whereas the REV model places its focus on the ethical *principles* that may be violated by COAs. This shift in perspective is intended to more adequately capture the ethical space surrounding a given situation so that, ultimately, the model can facilitate contextually-sensitive ethical decision-making.

Section 2 of this article presents brief background information on machine ethics and related attempts to define, quantify, and measure ethical harm, and then summarizes the authors' earlier work on quantifying the ethical consequences of military courses of action. Section 3 describes the design of the REV model and the ethical principles it is based on. Section 4 discusses the scenarios and survey used to collect the ethical assessments of expert and non-expert humans for military scenarios. Section 5 explains the process used to calibrate the REV model and identifies the statistics used to measure agreement in ethical assessments. Section 6 reports the results of comparing the REV model's ethical assessments with those of human experts. Finally, Section 7 presents the findings of this research and lists possible future work.

2. Background

This section presents a brief assessment on various tools and scales that address the ethical domain. First, the nature of defining ''evil,'' as well as precedents and inherent problems with doing so, are discussed. Then, a brief survey of research on machine ethics, some specific tools, and lessons learned from this domain are presented. Finally, the authors' earlier work on quantifying the ethical consequences military courses of action is summarized.

2.1 Defining and measuring "evil" and ethical harm

Psychologists have attempted to understand evil through scientific quantification and classification. Zimbardo describes the intent of Milgram's notorious studies of obedience as ''a paradigm in which it was possible to quantify 'evil' by the extremity of buttons pushed on a shock generator that allegedly delivered shocks to a mild-mannered confederate who played the role of the pupil or learner while the subject enacted the teacher role.''[1,2]

Baron-Cohen equates evil with a lack of empathy. He then focuses on clarifying what is meant by empathy as a means to contrast it with evil.[3] His definition requires both recognizing the other's thoughts and feelings as well as responding appropriately. His Empathy Quotient questionnaire classifies individuals in terms of a sliding scale, divided into levels of empathy ranging from zero empathy to empathic ''hyperarousal.''[4]

Welner's Depravity Scale has received considerable attention and support since 2001, but also some skepticism and criticism.[5,6] Welner's scale uses a mixture of subjective classifications, such as ''heinous'' and ''atrocious,'' and ''grave risk[s]'' to others, as well as objective classifications, such as a death to multiple victims, the use of a weapon, and property damage. Welner attempts to define subjective terms more precisely, based upon evidence and public survey, which is important because courts use such terms in determining appropriate sentencing for crimes.[7] Another example is Michael Stone's 22-level Scale of Evil, which is loosely based on *Dante's Inferno* and ranges from justified self-defense to ''schemers'' to ''psychopathic torture-murderers.''[8]

Evil is a very abstract phenomenon of a philosophical and psychological nature. The definitions of evil used in models such as the Empathy quotient, the Depravity Scale, the Scale of Evil, and the authors' earlier MOE model (explained later), are arguably too specific. Along with other factors, this led to a reframing of what was modeled—from evil to ethical violation—in the REV model.

Like evil, ethics (and thereby ethical violation), is difficult to define with both specificity and satisfaction. However, ethics is easier to pinpoint for the purposes of this research. The term ''ethics'' is derived from the Greek term *ethos*, or character.[9] Aristotle's most influential work on ethics, *Nicomachean Ethics*, ultimately states that ethics is concerned with *actions* that focus the larger community on ''the best good,'' and that, because of its scope, it is ''a sort of political science.''[9] To him, this ''best good'' is essentially happiness or general welfare. Moreover, in ethics, ''knowledge of [natural] principles provides us with'' this ''good to be pursued.''

This offers two key points on the nature of ethics and evil. First, ethics is explicitly focused on action. Second, it

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

can be decomposed into principles. Thus, while evil and ethics are both associated with a sense of right intent and character, ethics is more readily applicable than evil for evaluating actions in this research.

2.2 Machine ethics approaches

Traditional decision support tools and methods, especially those rooted in game theory, tend to rely heavily upon strategy and rational cause-and-effect to determine appropriate courses of action. They saw heavy use in military operations, most notably during World War II and the Cold War.[10] During the Cold War, for example, they drove the mutual assured destruction (MAD) doctrine, the conclusion that the United States and the Soviet Union would face annihilation if one were to launch a full-scale nuclear assault on another—which led to nuclear deterrence on both sides.

The nascent field of machine ethics involves the design and development of decision support models that facilitate *ethical* decision-making. Typically, these models use sets of rules, attributes, consequences, or principles associated with potential courses of action to provide suggestions, to guide the user through the ethical decision-making process, or to suggest an ethically sound action.

Two general perspectives on the nature of ethics drive tools based upon machine ethics: the consequentialist perspective and the principle-based perspective.[11] Consequentialist approaches for ethical reasoning are structured as weighted summations of ethical pleasure or goodness upon all individual people. Principle-based theories involve the use of universal laws, virtues, and intentions rather than direct consequences. As noted by McLaren, there is no agreement on which approach is best.[12] Machine ethicists largely deem consequentialist ethical interpretations as simple to implement but incomplete, and principle-based interpretations as a better match for human intuition but more complicated to implement and prone to an undesired level of subjectivity.[11–13] McLaren's dilemma has been stated as ''any tool that provide ethical judgments must necessarily 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.''[14] This implies that striking a prudent balance between objectivity and subjectivity is necessary for any tool dealing with the ethical domain.

2.3 Consequences-based evaluation of COAs: the metric of evil

The REV model, which is the subject of this article, was inspired by an earlier model known as the Metric of Evil (MOE), developed by the US Army and the University of Alabama in Huntsville.[15] As with the REV model, the earlier MOE model assessed military COAs from an ethical perspective and explicitly suggested the ''lesser of two evils'' of a pair of actions. True to its name, the MOE model explicitly attempted to measure ''evil,'' using ''intentional or anticipatable harm'' as a working definition of evil, as inspired by Zimbardo's own definition.[2] Its measurement of evil was based on quantified tangible consequences resulting from the COAs, weighted by relative ethical significance. Starting from the assumption that experts in the subject domain (in this case, military affairs) and in the ethical domain are the best possible ''objective'' measure of the potential harm associated with a COA, the parameters associated with the MOE model were calibrated to match these experts' ethical assessments of the COAs selected in real, historical military scenarios, such as the Warsaw Uprising (1944) and the Bay of Pigs invasion (1961). The optimal parameters for the MOE model produced good agreement between the model and expert assessments, which suggested potential for the concept behind the MOE model.[15]

However, the results of the study also raised significant questions about the MOE model. When optimized, the number of civilian casualties far outweighed every other quantifiable consequence in the MOE's calculations. Moreover, the calibration process produced a value for another model parameter associated with the count of a consequence such that the presence of a non-zero value for any given consequence mattered much more than the number of instances of that consequence. Taken together, this suggested that the only measure that experts used to evaluate the ''evil'' or harm associated with the COA was whether or not causing civilian casualties was a direct intent of the COA.

This result may align well with one's intuition on ethical matters—that if an action involves the intentional killing of innocents, then it is considered to be evil, virtually regardless of any other consequences of the action. However, the results also suggested that the MOE model's approach does not adequately parameterize the full range of ethical factors that humans actually use in making ethical assessments.[15] The research indicated that the MOE model could not readily weigh the loss of human life against the promotion of ideals or principles. This suggested that a different analysis of the ethical principles applicable to military COAs would be more meaningful and relevant for evaluating actions. One comment submitted by a participant in the MOE experiment illuminates an important component that was missing from the MOE model:

> How many casualties are justified in the promotion or defense of capitalism, democracy, communism, fundamentalist Islam, or the power of a warlord?[15]

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

These questions regarding the MOE model suggested that a new model, one based on a different means of ethical assessment of COAs, was needed.

3. REV model

This section describes the design and mathematical structure of the REV model and then describes the process of selecting the principles the model is based on.

3.1 Model design

Some existing automated ethical reasoning programs, such as those developed by McLaren,[12] use sophisticated artificial intelligence techniques at their core. However, there can be significant power in simple modeling approaches, especially when tied to the psychological realm. In support of this idea, Dawes found that, in making decisions, people excel in determining important factors but not in integrating information.[16] She describes simple linear models constructed upon just a few key factors, such as a predictive model of faculty ratings of students based solely upon their grade point averages and Graduate Record Exam (GRE) scores,[17] that are quite powerful. In fact, Dawes's studies have shown that even models whose weights have been chosen *randomly* have outperformed human judges, so long as appropriate inputs are chosen for the model. She concluded that ''[t]he whole trick is to decide what variables to look at and then know how to add.''[18]

With Dawes' findings in mind, it is less surprising that many decision support tools and decision analysis techniques make use of linear models.[19] The key differences between these tools are often in their respective weighting schemes. Simple linear models can produce significant results, so long as a proper set of inputs is chosen.

The REV model uses a linear model to perform an ethical comparison between two COAs. Its internal calculations consider violations of ethical principles, with violations of multiple principles considered separately, and then combined. Formally, given two alternative military COAs denoted A and B, the REV model evaluates their relative ethical violation as the quantity ΔV_{AB}, calculated as

$$\Delta V_{AB} = \sum_{j=1}^{n_p} (v_{Bj} - v_{Aj}) w_j \qquad (1)$$

where n_p is the number of ethical principles considered, $w_1, w_2, \ldots, w_{n_p}$ are the relative weights of those principles, and v_{Aj} and v_{Bj} are the extent to which COAs A and B, respectively, violate a given principle j. Essentially, w_j is a measure of how important adherence to a given principle is to avoiding overall ethical violation, relative to other principles. The REV model then suggests COA A if the resulting ΔV_{AB} is less than 0, COA B if it is greater than 0, and neither otherwise.

The mathematics of the REV model as given by equation (1) may seem unexpectedly simple, but as Dawes found, such models can be quite powerful. The power of the model stems from the proper selection of a set of principles relevant to military COAs and in correctly calibrating the weights of these principles.

3.2 Choosing a set of principles

For the REV model to compare COAs based on their violation of ethical principles, a suitable set of principles for it to use must be identified. The specific principles used by the REV model were extracted from four relevant and authoritative sources in the literature: Ross's *prima facie* duties, a set of four biomedical ethical principles, the United States Law of Armed Conflict, and Just War theory.

Ross's set of *prima facie* duties is a *general* set of ethical principles that incorporates a broad philosophical perspective.[20] He enumerates them and describes their context in detail. He lists seven principles (six explicitly, with one general principle decomposed into two). These include fidelity, couched in terms of holding true to a ''promise or […] implicit promise;'' reparation for previous ''wrongful act[s];''''duties of gratitude'' that ''rest on previous [services] of other men;''''duties of justice'' that are concerned with ''distribution of pleasure or happiness'' based upon individual merit; ''duties of self-improvement,'' such as ''virtue or of intelligence;'' and ''non-maleficence […] as a duty distinct from that of beneficence.''

From Ross's point of view, ''there is nothing arbitrary'' about these principles, and humanity ''knows [our main convictions] to be true.'' He does, however, list his principles ''without claiming completeness or finality,'' also realizing that he ''certainly cannot prove to [readers]'' that, for example, to make a promise to another is ''to create a moral claim.'' Despite these limitations, this set does capture and consolidates many ethical principles used in other contexts, including that of the US military and of vastly differing schools of philosophical thought.[21–23]

Another set used for this research encompasses four principles for biomedical ethics. Originally presented by Beauchamp and Childress,[24] and heavily championed by Gillon,[25,26] the four comprise a notable set of principles in widespread use in the field. The set is best summarized as ''respect for autonomy, beneficence, non-maleficence, and justice.''[25] Gillon describes them as ''*prima facie* principles'' in an explicit reference to Ross's use of the term. Two of these principles, by nomenclature alone, clearly overlap with Ross's set—beneficence and non-maleficence—which Gillon describes as invoking the Hippocratic moral obligation.[25]

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

Like the military domain, the biomedical domain is one in which ethics, by necessity, has been considered in depth. Gillon claims that these four principles can ''explain and justify, alone or in combination, all the substantive and universalisable claims of medical ethics and probably of ethics more generally''.[26] To him, most notable is autonomy, or ''self-rule,'' which Gillon asserts is the ''first among equals.'' In essence, this principle requires physicians ''to obtain informed consent from patients before we do things to try to help them.'' Ultimately, then, physicians are in large part bound to the authority of the patient when administering treatment.

In addition to the above principle sets, certain individual countries maintain Laws of Armed Conflict (LOAC) guides,[27,28] from which sets of ethical principles directly related to military conflict can be derived. Many aspects of specific countries' LOAC guides generally overlap. Principles for a third set for comparison and analysis were extracted from the United States LOAC.[27]

While the LOAC guide is extensive and detailed, several key principles on methods and reasons for warfare are readily apparent. Most emphatically, the LOAC distinguishes between civilian and military persons and assets, stating that ''[c]ivilians and civilian property may not be the subject or sole object of a military attack'' and that ''[o]nly combatants or those directly participating in hostilities may be targeted.'' It also prohibits ''attack[s] […] which would be excessive in relation to the concrete and direct military,'' clearly stating that this ''violates the principle of proportionality.'' The LOAC also states that ''[f]orce should be used as a 'last resort.'''

In addition, the LOAC states that ''all uses of force [require] both the necessity and proportionality criteria,'' and only those of ''legitimate authority (those who rule, i.e., the sovereign)'' have the right to make decisions on military action—just as Gillon claims that a patient, rather than the physician, has authority over whether or not the physician is authorized to take action. In addition to these, the LOAC describes another historically relevant ethical principle, independent of the above: that there must be a reasonable ''prospect of victory,'' or ''[p]robability of success.''

Finally, the Just War theory, with its centuries-long history and a large number of adaptations, is a primary driver for other sets of principles—including the United States LOAC.[27] The Catholic Church's teachings of the theory have been one of the most influential, and its Catechism captures the Church's perspective on the matter.[29] The Catechism's section on ''Safeguarding Peace'' was seen as most appropriate for this research. It discusses more abstract concepts such as anger, hatred, and revenge, but it also provides concrete principles.

Its most apparent principles are explicitly listed in its paragraph 2309 as ''the traditional elements […] in what is called the 'just war' doctrine''—what it refers to as the ''strict conditions for legitimate defense by military force.'' The Catechism declares that other means to establish peace with an aggressor who has inflicted ''lasting, grave, and certain'' damage ''must have been shown to be impractical or ineffective,'' that the action must have ''serious prospects of success''; and that the action ''must not produce evils and disorders graver than the evil to be eliminated.'' That is, an action under consideration must not cause disproportionate damage, must be absolutely necessary, and hold a serious probability of success. Elsewhere, the text explains that ''non-combatants, wounded soldiers, and prisoners must be respected and treated humanely'' and that ''every act of war directed to the indiscriminate destruction of whole cities or vast areas with their inhabitants is a crime''; thus, discriminating between those who are partaking directly in the conflict is a necessity, especially when a large number of civilian lives are at stake.

Ultimately, the Catechism claims that ''[i]njustice, excessive economic or social inequalities, envy, distrust, and pride raging among men and nations threaten peace and cause wars.'' It further claims that actions taken to ''overcome these disorders'' are intended to ''build […] up peace and avoid […] war,'' akin to Aristotle's claim that general welfare is an end in and of itself.[9] The Catechism maintains that human reason contributes in ''assert[ing] the permanent validity of the moral law during armed conflict''; war does not necessitate that ''everything becomes licit between the warring parties.''

Table 1 summarizes a clustering of principles present in these source principle sets. The sets were examined to identify principles that were equivalent (or at least very similar) across sets. Each row in the table corresponds to a principle found in one or more of the principle sets used as sources. The table also provides the term that each set uses for every principle—for example, the Catechism's Just War doctrine mentions discriminate destruction along with proportionality, both contributing to its assessment of the general principle of proportionality. If a principle is not present in a given set, the corresponding cell in the table is empty. (It is also worth noting that principle sets that were not chosen as source sets also contained several of these same principles.)

Each of the source principle sets listed in the table header was assigned a selection weight. (These selection weights were used to select principles for the REV model; they are not the weights used in the model's calculations.) Ross's *prima facie* and the set of medical principles are assigned selection weights of 1 (lower), whereas the LOAC and Just War principles are assigned selection weights of 2 (higher); the latter's ratings are higher because they specifically involve ethics in the military domain. Each principle's score is the sum of the selection

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

Table 1. Analysis of ethical principles derived from the literature. Each of the principles found in the source principle sets, their names or representations of the principle sets, and their score in the selection process are shown.

Principle	Principle set and selection weight				Score
	Prima facie (weight 1)	Medical (weight 1)	LOAC (weight 2)	Just War (weight 2)	
Civilian non-maleficence	Non-maleficence	Non-maleficence	Distinction	Discrimination	6
Proportionality	Justice		Proportionality	Proportionality, Discriminate destruction	5
Necessity			Necessity	Last resort, Avoiding war, Response to severe damage	4
Prospect of success			Probability of success	Serious prospects of success	4
Sole power to authorize action		Autonomy	Proper authority		3
Beneficence	Beneficence	Beneficence			2
Combatant non-maleficence	Non-maleficence	Non-maleficence			2
Fidelity	Fidelity				1
Self-improvement	Self-improvement				1
Reparation	Reparation				1
Gratitude	Gratitude				1

weights of the source principle sets it appears in. Thus, principles that appear in more of the source sets, and in more heavily weighted sets, will have higher selection scores. The maximum possible score is 6. Principles with scores greater than half of the maximum score (that is, 4 or greater) were selected for the REV model. In order of conjectured importance, the ethical principles chosen for the REV model are the following:

- *Civilian non-maleficence* (p_1). This principle requires conducting military actions so as avoid harm—especially intentional harm—to civilians. This principle is most related to the result of the earlier MOE research; since, when calibrated, the MOE model considered intentional harm to civilians as the overwhelmingly dominant ethical factor.
- *Necessity* (p_2). This principle requires that a military action be militarily necessary and that other attempts for peaceful resolution have not been fruitful.
- *Proportionality* (p_3). Because the goal of military action is said to be restore peace with an aggressor[29], this principle requires that a military action not cause damage disproportionately in excess of that caused by the aggression.
- *Prospect of success* (p_4). This principle requires that a military action should not inflict harm for a ''lost cause'', i.e., the action should have a reasonable chance of succeeding to justify any casualties and destruction it may cause.

4. Data collection

Calibrating and evaluating the effectiveness of the REV model required comparing its assessments of ethical violations to those of human experts for realistic military scenarios. This section discusses the scenarios used to calibrate the REV model and the survey process used to collect human assessments of those scenarios.

4.1 Calibration scenarios

As a basis for calibrating the REV model, realistic military scenarios that involved violation of ethical principles were developed. As discussed earlier, four principles were selected as inputs to the REV model. Each calibration scenario was designed to contrapose, or force a choice between, two of the four selected principles, and to neutralize or leave out the other two principles as much as possible while maintaining scenario realism. Each of a scenario's two COAs violated one or the other, but not both, of the two ethical principles the scenario was designed for. For example, Scenario 1 has two COAs designed so that one COA violates principle p_1 (*civilian non-maleficence*), the other COA violates principle p_2 (*necessity*), and neither COA violates p_3 (*proportionality*) or p_4 (*prospect of success*) in any significant way. Given four principles, there are six possible pairs of principles (order is not significant and the principles were not paired against themselves); hence six scenarios were developed, one for each pair of principles.

The six calibration scenarios were notional. The use of notional scenarios offered flexibility in developing

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

tradeoffs between ethical principles and reduced the potential for cultural and affiliatory biases among humans assessing them. Although the scenarios are notional, to maintain realism they were based on actual historical events, and to maintain relevance they involve modern military operations. All of the historical events the scenarios are based on took place during or after World War II, and two of them were ongoing as of September 2014. The scenarios are drawn from both conventional and counterinsurgency warfare, and include land, sea, and air operations.

For each scenario, a survey participant was provided with the scenario title, a brief narrative description of the situation, and statements of two alternative COAs (internally designated as COA *A* and *B*, respectively). The participant's assessment of the two COAs' relative ethical violation was elicited via a Likert scale,[30] with the choices ranging from one COA being clearly ethically preferable through intermediate assessments to the other COA being clearly ethically preferable. The scenario descriptions also specify numeric values for the tangible consequences, such as estimated civilian casualties, of the COAs. These provide background information for the human survey participants and were used later as input to the MOE model, which requires numerical consequences as input, when the performance of the MOE model was compared to the REV model.

Also associated with each scenario, but hidden from the participants, is an identification of the principle violated by each of the scenario's COAs and a measure of intent associated with each of its tangible consequences. The REV model and MOE model respectively use these additional inputs to perform their assessments. Table 2 summarizes the six scenarios.

To illustrate the calibration scenarios and demonstrate how they were designed to force a choice between two ethical principles, the full text for Scenario 2 ''The border encroachment'' follows. The scenario's ground assault COA violates the civilian non-maleficence principle (p_1) and the air strike COA violates the proportionality principle (p_3). The Likert scale choices available to the survey participant are at the end of the example. The country names in the scenarios are always given as randomly selected colors when a participant is given the survey.

Scenario 2. The border encroachment

You are a military decision-maker for the nation of ***Green****, engaged in conflict against* ***Purple****.*

Green is engaged in an ongoing series of conventional, low-intensity military skirmishes with Purple over a disputed province that lies on their shared border. The province is currently controlled by Green, but the civilian population of the province is an intermingled mixture of ethnic and religious groups, some of whom identify with Green and others with Purple. Throughout the three-year period of these skirmishes, Purple and Green have each suffered around 2500 combatant casualties. Both Purple and Green have a small arsenal of moderate-yield nuclear weapons; so far neither nation has used them in any conflict, but the resulting threat of nuclear exchange creates great tension between the two nations, intensifying the stakes behind this border dispute.

Recently a medium-sized force of Purple combatants unexpectedly crossed the border and occupied a village within the disputed province, quickly overwhelming the small detachment of Green soldiers manning an observation post nearby. The citizens of the village, which include both Green and Purple sympathizers, were not allowed to leave the village and are now essentially prisoners or hostages. During the Purple attack and occupation one Green historical cultural site and one facility that provided essential utilities to surrounding civilian population were destroyed.

The Green military leadership has identified two courses of action that are certain to force Purple to retreat from the village and reestablish Green control. You must decide which course of action to use to recapture the village.

Use a ground assault

A direct ground assault on the village with a large Green force will drive the Purple combatants out of the village. However, the fighting within the village will endanger the village's population and risk the destruction of facilities and buildings. Because Green can achieve quick victory by deploying an overwhelming force, combatant casualties are expected to be comparatively light. In the operation, Green will face 25 combatant casualties, while Purple will face around 40 military casualties.

The ground assault is estimated to result in consequential civilian casualties of around 350 Green citizens and around 350 Purple (but not hostile) civilian immigrants to the village. The operation must also necessarily destroy two Green cultural sites that have been seized by Purple combatants.

Use an air strike

An air strike against a major Purple military base located just over the border would also force the occupying Purple troops to retreat from the contested village because their logistical support comes from that base. However, the Purple base has extensive air defenses, including radar sites, anti-aircraft guns, and surface-to-air missiles. Consequently, a successful Green air strike will require a large strike force and will likely result in many combatant casualties for the Purple defenders. Moreover, the destruction of the important Purple base is likely to be seen by Purple as an intentional and dangerous escalation of the conflict.

The air strike is expected to cause 5000 combatant casualties, mostly Purple. It is also expected to damage or destroy 10 historical landmarks and 5 facilities that provide essential utilities to the Purple population located near the military base. In anticipation of this or other potential Green responses, however, Purple has evacuated its civilians from the area near the base and therefore will not suffer any civilian casualties.

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

Table 2. A summary of the scenarios used in the calibration process, including each scenario's two COAs and the principle that each COA violates.

Scenario 1. Intercontinental warfare	
Historical basis: US submarine campaign against Japan, World War II, 1945	
COA	**Principle violated by the COA**
A. Discontinuing a submarine blockade	p_1. Civilian non-maleficence
B. Continuing a submarine blockade	p_2. Necessity
Scenario 2. The border encroachment	
Historical basis: India–Pakistan dispute over Kashmir province, ongoing	
COA	**Principle violated by the COA**
A. Using a ground assault	p_1. Civilian non-maleficence
B. Using an air strike	p_3. Proportionality
Scenario 3. Stabilization	
Historical basis: US operations in Afghanistan, ongoing	
COA	**Principle violated by the COA**
A. Launching an urban sweep	p_1. Civilian non-maleficence
B. Continuing rural patrols	p_4. Prospect of success
Scenario 4. The archipelago	
Historical basis: Sinking of the ARA *General Belgrano*, Falklands War, 1982	
COA	**Principle violated by the COA**
A. Sinking an escorting warship	p_2. Necessity
B. Sinking a transport	p_3. Proportionality
Scenario 5. Last-ditch offensive	
Historical basis: Battle of the Bulge, World War II, 1944–1945	
COA	**Principle violated by the COA**
A. Surrendering unconditionally	p_2. Necessity
B. Launching a counter-offensive	p_4. Prospect of success
Scenario 6. Merchant ship crew recovery	
Historical basis: Seizure of the SS *Mayaguez*, Southeast Asia, 1975	
COA	**Principle violated by the COA**
A. Launching a military recovery operation	p_3. Proportionality
B. Seizing a ship and negotiating	p_4. Prospect of success

Decision
Given only the information above, is ***using a ground assault*** *or* ***using an air strike*** *ethically preferable?*

- *Using a ground assault is clearly ethically preferable.*
- *Using a ground assault is somewhat ethically preferable.*
- *Neither action is ethically preferable over the other.*
- *Using an air strike is somewhat ethically preferable.*
- *Using an air strike is clearly ethically preferable.*

4.2 Survey process

An online survey was used to collect the participants' responses to the scenarios. Supporting software was implemented using the PHP scripting language to generate survey instances for each participant and the MySQL database system to store participant responses. Survey instances were stored with a unique identification number associated with them, but they were never linked to the identity of a specific participant; the completed surveys were fully anonymous, with the identities of the participants unknown even to the experimenters.

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

When a participant first visited the survey website, he or she was greeted with an introductory page that provided information on the project. When the participant proceeded, the software generated a survey instance for the participant. This survey instance presented all six scenarios to the participant. The software randomized several aspects of the presentation, which mitigated unintended biases. First, the scenarios themselves were presented in random order. Secondly, the two potential COAs within each scenario were presented in random order (and the choices in each decision block followed this same order). Third, names for countries involved in the scenarios were randomly chosen among a variety of colors; a country may be named Green, Orange, or Purple, for example, but Blue and Red were excluded because of potentially biasing connotations these colors have acquired through repeated usage. The mechanism ensured that no country name was used in more than one scenario, which conveyed the notion that the scenarios presented to the participant were intended to be independent from one another.

After the participant chose a response to each scenario, he or she was prompted to self-classify his or her expertise, encoded as a forced choice between military expertise, humanities expertise, or neither. Descriptions of each choice were provided, but the explicit terms ''military expertise'' and ''humanities expertise'' were hidden from the participant. These terms were not used in the descriptions themselves in order to force the participant to use relatively specific criteria to evaluate their expertise. The ordering of the military and humanities expertise descriptions was also randomized to mitigate bias toward one or the other (though that of ''non-expertise'' was always presented last). Descriptions associated with military expertise, humanities expertise, and non-expertise, respectively, were provided as follows:

- I have specific training or experience, gained in military service; or I am an active, retired, military civilian, government civilian, or reserve component personnel.
- I hold an advanced degree in philosophy, psychology, political science, sociology, history, ethics, humanities, or a related field; or I hold a position as a religious leader or another similar position; or I hold a position as a counselor, social worker, clinical psychologist, or in other related work.
- My expertise/experience is not described by either of the above.

Once the participant's scenario assessments and self-classification were collected, the mechanism recorded the amount of time that the participant spent on the scenario assessment page (which excluded the introductory page and subsequent pages). The website then presented the participant with the option to either close the website or to provide additional optional demographic information. The optimal information included specific education, military rank (if applicable), gender, and age. Prompts for each piece of multiple-choice information included ''I don't want to specify,'' and those of free-form information were left blank by default. This ensured that participants were able to provide only the information that they wished to.

The online survey mechanism was subjected to a pilot test that spanned two weeks. This test was intended to ensure that the website mechanism functioned correctly, that the scenarios were comprehensible, and that participants had an appropriate cognitive load—that is, that they spent neither too little nor too much time in making their assessments. Overall, the pilot test received 97 responses. The data collected indicated that the mechanism was correctly displaying and gathering information. It further suggested that the scenarios provided an adequate cognitive load, as participants typically spent around 10–15 minutes assessing the entire collection of scenarios. Participant responses collected in this pilot were used only to perform this test; they were not used in the actual experiment.

For the experiment, human participants were recruited from a variety of sources. Recruitment of non-expert participants from local and non-local sources, including from outside of the United States, was straightforward. To secure expert assessments, the investigators directly contacted professional and personal colleagues known to have expertise related to the military and/or humanities. Persons at nearby Redstone Arsenal and within the University of Alabama in Huntsville College of Liberal Arts faculty were also contacted. In additional to inviting them to participate in the survey, these experts were requested to forward the survey to other experts—many of them had their own set of contacts from which to draw more experts. The search pool was then extended to faculty and contacts at other universities within the United States, especially those with strong philosophy departments and organizations or centers focused on ethics. Many universities outside of the United States were also contacted. In total, more than 1000 people were directly contacted, with many others indirectly contacted.

A total of 141 human participants provided responses. Of those that responded, 78% also provided at least some optional demographic information. The survey data showed no relationship between participants' ages and valued principles. Moreover, 18% of those who provided some demographic information specified that they have had combat experience, the average age for those who disclosed it was 47 years, 7% of respondents lived outside of the United States, and the average time for participants to assess all scenarios was 17 minutes, 34 seconds. Regarding formal education, 29% reported that they had bachelor's degrees, 35% master's degrees, and 18% doctoral degree,

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

with each roughly evenly split between military experts, humanities experts, and non-experts.

5. Model calibration

This section explains the process used to calibrate the REV model, i.e., to find the best values for the model's principle weights. It defines the different classes of humans and models compared, details the calibration process, identifies the statistics used to measure agreement, describes two additional variants of the REV model that were included in the comparisons, and reports the principle weights found during the calibration process.

5.1 Raters

The standard of comparison for the model calibration process was the collective assessment of human experts, drawn from both the military and the humanities, on the relative ethical violation of the scenarios' COAs. We base the quantitative comparison of scenario assessments on the notion of a rater. A *rater* is any human, model, or process that assesses the scenarios for ethical violation. Three classes of human raters were used to calibrate and evaluate the REV model: military experts, humanities experts, and non-experts. Military experts (51 human raters) and humanities experts (35 human raters) provided the standard for comparison when assessing the performance of the models. Raters with military expertise were categorized as experts because they have specific training and experience in military decision making, including COA selection. Raters with expertise in humanities were categorized as experts because their training and experience in critical thinking and comparative analysis would allow them to make informed, well-considered, and supportable ethical decisions from a perspective that was not military. Military experts, even if asked to select a COA from an ethical perspective only, may in some cases be unintentionally influenced by their estimates of the COAs' military effectiveness. Non-experts (55 raters) were included for comparison with the expert human raters.

A class of computer generated random raters (1000 raters), that simply selected one of the five responses randomly, were included to provide a minimal performance threshold. The models would have to outperform random raters in terms of agreement with the experts to be considered to have any utility at all.

The calibrated REV model and recalibrated MOE model were each a class with one rater. Moreover, two additional variants of the REV model, differing only in the principle weights used by the model, were defined and treated as rater classes; these will be explained later.

Each rater belonged to exactly one of these eight base classes. In addition to these base classes, two composite classes of raters were considered: all experts and all humans. The class of all experts included both military experts and humanities experts (51 + 35 = 86 raters), and the class of all humans includes both classes of expert humans as well as non-expert humans (51 + 35 + 55 = 141 raters).

5.2 Model calibration process

Essentially, the REV model was calibrated by iteratively setting values for the principle weights w_1, w_2, w_3, w_4 and comparing the model's assessments of the scenarios' COAs using those weights to the assessments of the expert raters. (A more precise statement of the procedure used for each iteration will follow after some preparatory explanation.) The goal of the overall calibration process was to find values for the weights that would gave the best possible agreement of the REV model with the expert raters. The MOE model was also calibrated so that the REV model and MOE model approaches could be compared directly.

During the calibration process, numeric measures of agreement between individual raters and classes of raters were used. The numeric agreement between any two individual raters was calculated from the number of scenario assessments for which the two raters agreed and disagreed. Two raters' assessments of a particular scenario were said to agree if they both considered the same COA to be ethically preferable to the alternative COA, regardless of whether it was *somewhat* or *clearly* ethically preferable, or if they both found that neither COA was ethically preferable. For example, if for a given scenario raters R_1 and R_2 both assessed COA A to be ''clearly ethically preferable,'' then the were considered to agree, whereas if R_1 assessed COA A as ''clearly ethically preferable'' and R_2 assessed COA B as ''somewhat ethically preferable,'' then they were considered to disagree. For any given pair of raters, each of the first rater's six scenario assessments was compared to the assessment of the same scenario made by the other rater.

The agreement between two classes of raters was determined by comparing every rater in the first class to every rater in the other class in the manner just described. The agreement between two rater classes was quantified and normalized for different rater class sizes using a statistical measure of agreement, Goodman and Kruskal's Γ, which will be explained in the next section.

In each iteration of the calibration process, the agreement of each class of raters with the expert raters was measured. A simple approach to measuring the rater classes' agreement with the experts would have been to compare each of the other classes with the class of all experts. However, the authors wanted to measure agreement among the experts as well. To accomplish that, in

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

each calibration iteration the class of all experts was randomly partitioned into two subsets of equal size. One of the two subsets of the class of all experts was treated as the comparison set, against whom the other classes' agreement was measured; that subset was called the ''standard'' raters. The other subset of the class of all experts was treated as simply another class of raters to be compared to the standard raters; that subset of the experts was included in the ''contestant'' classes of raters, along with all of the other rater classes (non-experts, random, the REV model and its two variants, and MOE model). The partitioning of the expert raters into standard and contestant raters made it possible to measure the agreement of the expert raters with other expert raters, and introduced the possibility of finding that expert raters did not agree well among themselves. To ensure a single specific random partition of the experts into standard and contestant raters did not unduly influence the agreement measures, the random partitioning of the class of all experts into contestants and standards and the calculation of the agreement measures between each of the contestant classes and the standards was repeated at least 300 times for each calibration iteration, i.e., for each set of REV model principle weights tested.

The calibration procedure was implemented in software (specifically PHP), which simplified the execution of multiple iterations and allowed automatic access to the collected survey data. Each iteration of the calibration process used the following procedure:

1. Generate a set of REV model weights. (How this was done will be described later.)
2. Execute the REV model and its two variants to generate scenario assessments.
3. Execute the MOE model to generate scenario assessments.
4. Generate random scenario assessments for each of the random raters.
5. Repeat either 300 or 1000 times:
   - 5.1 Randomly partition the class of all experts into two equal size subsets, denoted *standard* raters and *contestant* raters. The standard raters will serve as the standard for comparison for all other raters.
   - 5.2 Add the raters in all of the other rater classes (non-experts, random, REV model and its two variants, and the MOE model) to the set of experts selected as contestant raters.
   - 5.3 Repeat once for each contestant rater:
     - 5.3.1. Compare the scenario assessments of the contestant rater to those of each of the standard raters for each of the six scenarios, counting the number of times the contestant rater agrees or disagrees with each of the standard raters on a scenario.
     - 5.3.2. From those counts, calculate and save a numeric value quantifying the contestant rater's agreement with the class of standard raters.
6. For each of the six classes of contestant raters, calculate the mean of the agreement values found in step 5.3.2 for each member of that class.

The calibration process proceeded in two phases. In the first phase, the parameter space of the principle weights was methodically searched. The calibration procedure was executed for every possible combination of the five values 0.00, 0.25, 0.50, 0.75, and 1.00 for each of the four principle weights w_1, w_2, w_3, and w_4. During those $5^4 = 625$ iterations, the partitioning of the class of all experts in step 5.1 was repeated 300 times per iteration. The best set of principle weights found in the first phase was then brought into the second phase. In the second phase, a Monte Carlo process was used to randomly generate small variations of the best principle weights, and each of those variations was used for an iteration of the calibration procedure. During the second phase, the partitioning of the class of all experts was repeated 1000 times per iteration.

In total, approximately 100 iterations of the calibration procedure were executed in the second phase. The calibration iterations were terminated when an effective set of REV model weights had been found and the incremental improvements to the REV model's agreement with the experts, from one iteration to the next, were less than 10^{-4}.

5.3 Agreement statistics

A statistical measure of agreement, Goodman and Kruskal's Γ statistic,[31] was used to measure the agreement between classes of raters. Other statistics sometimes used for similar applications include Pearson's τ, Krippendorf's α, and Kendall's τ series (which is not related to Pearson's τ).[32] Goodman and Kruskal's Γ statistic is suitable when the data is ordinal, not nominal, which was the case in this analysis; e.g., while ''clearly ethically preferable'' in a scenario's Likert responses is certainly more preferable than ''somewhat ethically preferable'', it cannot be assumed to be two (or any other constant) times as preferable. In general, the Γ statistic is calculated as:

$$\Gamma = \frac{C - D}{C + D} \tag{2}$$

where C is the number of concordant pairs (in this case the number of agreements between two raters on a scenario) and D the number of discordant pairs (in this case the number of disagreements between two raters on a scenario). The possible values of Γ range from -1 to 1, with

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

positive values indicating agreement, negative values indicating disagreement, and values near 0 indicating about as much agreement as disagreement.

To compare the performance of the rater classes, it would have been ideal to measure the cumulative agreement of the individual raters in each contestant class with the individual standard raters for all possible partitions of all experts into standard and contestant raters. However, there are $C(86, 43) \asymp 6.64 \times 10^{24}$ possible partitions, and thus calculating agreement for all possible partitions is infeasible. Instead, the authors measured the agreement of the raters in each rater class with a randomly generated subset of the partitions, and then performed pairwise comparisons of the agreement for the rater classes using a conventional statistical hypothesis test. To explain this in more detail, we introduce some notation:

R A rater class; R_1 and R_2 denote any two different rater classes.
r An individual rater.
i A single partition of the class of All expert raters into standards and contestants.
$C_{r,i}$ Concordant pairs for individual rater r with all standard raters for partition i.
$D_{r,i}$ Discordant pairs for individual rater r with all standard raters for partition i.
$\Gamma_{r,i}$ Agreement of individual rater r with all standard raters for partition i.
$\Gamma_{R,i}$ Agreement of all raters in class R with all standard raters for partition i.
$\bar{\Gamma}_R$ Mean agreement of all raters in class R with all standard raters for all partitions in sample.

In step 5.3 of the calibration procedure, a $\Gamma_{r,i}$ value was calculated for each contestant rater from the total number of agreements (concordant pairs, C) and disagreements (discordant pairs, D) a contestant rater had with all of the standard raters for each of the six scenarios. Comparing a single rater as a contestant to all of the standard raters for all six scenarios for one partitioning of the expert raters into standards and contestants will be referred to as a single *observation*. In other words, one observation is one execution of step 5.3 in the calibration process for one rater in the class. One $\Gamma_{r,i}$ value is computed for each observation.

The agreement of a single rater with all standard raters for partition i would be calculated using

$$\Gamma_{r,i} = (C_{r,i} - D_{r,i})/(C_{r,i} + D_{r,i}) \tag{3}$$

and the agreement of all raters in class R with all standard raters for partition i with

$$\Gamma_{R,i} = \left(\sum_{r\in R} C_{r,i} - \sum_{r\in R} D_{r,i}\right) \Big/ \left(\sum_{r\in R} C_{r,i} + \sum_{r\in R} D_{r,i}\right) \tag{4}$$

Intuitively, $\bar{\Gamma}_R$, i.e., the mean agreement of all raters in class R with all standard raters for all 1000 partitions in the sample, would be calculated as:

$$\bar{\Gamma}_R = \frac{1}{1,000} \sum_{i=1}^{1,000} \left(\sum_{r\in R} C_{r,i} - \sum_{r\in R} D_{r,i}\right) \Big/ \left(\sum_{r\in R} C_{r,i} + \sum_{r\in R} D_{r,i}\right) \tag{5}$$

where the term inside the summation is the equation for $\Gamma_{R,i}$.

However, as described in the calibration procedure in the previous section, $\bar{\Gamma}_R$ for rater class R over the 1000 partitions was instead calculated as:

$$\bar{\Gamma}_R = \frac{1}{1,000 \cdot |R|} \sum_{i=1}^{1,000} \sum_{r\in R} (C_{r,i} - D_{r,i})/(C_{r,i} + D_{r,i}) \tag{6}$$

where the term inside the summation is the equation for $\Gamma_{r,i}$.

In the general case, the quotient of two summations is not equal to the summation of the corresponding quotients, and thus equations (5) and (6) cannot generally be assumed to be equal. However, in this particular case there is an additional constraint: $C_{r,i} + D_{r,i}$ has the same constant value (number of scenarios ã number of standard raters) for every rater r in a class R. Given that constraint it is easy to show that equations (5) and (6) are algebraically equivalent.

The $\bar{\Gamma}_R$ values calculated for each rater class measured the degree to which the raters of that class agreed with the expert raters. Given any two classes of raters R_1 and R_2 and their mean agreement values $\bar{\Gamma}_1$ and $\bar{\Gamma}_2$, the larger of $\bar{\Gamma}_1$ and $\bar{\Gamma}_2$ indicates the class with better agreement with the expert raters. To determine if the calculated differences in the $\bar{\Gamma}_R$ values for the different rater classes were statistically significant, pairwise comparisons of the $\bar{\Gamma}_R$ values for the rater classes using a conventional statistical hypothesis test were performed. In statistical terms, the populations for each test were the mean agreements of the raters in the two classes with the experts for all possible partitions, and the samples were the agreements of the raters in the two classes with the experts for the 1000 randomly generated partitions. Differences in the $\bar{\Gamma}_R$ values for two rater classes were tested for statistical significance using a conventional one-tailed hypothesis test for the difference of two means. Let μ_1 and μ_2 denote the unknown population means that would result from averaging the individual $\Gamma_{R,i}$ values for the raters in two rater classes R_1 and R_2 for all possible partitions of the expert raters into standard and contestant raters. $\bar{\Gamma}_1$ and $\bar{\Gamma}_2$ are the sample means calculated for classes R_1 and R_2 for the randomly generated

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

partitions, which are estimates of population means μ_1 and μ_2. The one-tailed hypothesis test has this form:

H_0: $\mu_1 = \mu_2$, i.e., the agreement of rater class R_1 with the class of all experts = the agreement of rater class R_2 with the class of all experts
H_1: $\mu_1 > \mu_2$, i.e., the agreement of rater class R_1 with the class of all experts > the agreement of rater class R_2 with the class of all experts

Because population standard deviations σ_1 and σ_2 were not known, they were estimated using the sample standard deviations s_1 and s_2, and the Student t distribution was used for the hypothesis test.[33] The sizes of the two samples were the same, $n_1 = n_2 = 1000$. The test statistic t was calculated as

$$t = \frac{\bar{\Gamma}_1 - \bar{\Gamma}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (7)$$

For any two rater classes R_1 and R_2, if the P-value associated with this statistic is less than the level of significance $\alpha = 0.05$, then H_0 was rejected, and it was concluded that R_1 statistically significantly outperformed R_2. Satterthwaite's formula was used to estimate the degrees of freedom when calculating the P-values.

5.4 Variants of the REV model

The calibration procedure found specific values for the REV model's principle weights w_1, w_2, w_3, w_4 that produced the best agreement with the experts; the REV model with those weights may be considered the ''calibrated'' variant of the REV model. Two secondary variants of the REV model, which differ from the calibrated variant only in the values for the weights, were also devised and used to test other interesting ideas.

The ''dominant p_1'' variant of the REV model made p_1 the only principle considered by the model by setting $w_1 = 1$ and $w_2 = w_3 = w_4 = 0$. Recall that principle p_1 is civilian non-maleficence and that the earlier MOE research suggested that intentional civilian harm overshadowed all other tangible factors in assessing ethical harm. Thus this REV model variant is the REV analog to the MOE model. The rating of this REV model variant would provide another perspective on the question of whether experts (and humans in general) weigh direct civilian harm over *all* factors—including intangible principles.

The ''all weights equal'' variant of the REV model had all weights set to the same value—$w_1 = w_2 = w_3 = w_4 = 1$. This variant tests an assumption implicitly made by several other ethical decision support tools and implied by many sets of ethical codes found in the literature—that all principles are to be considered equally.

5.5 Calibrated model weights

The calibration process for the REV model found that the following weights allowed it to best match the experts:

- For *Civilian non-maleficence* (p_1), $w_1 = 1.00$
- For *Necessity* (p_2), $w_2 = 0.65$
- For *Proportionality* (p_3), $w_3 = 0.25$
- For *Prospect of success* (p_4), $w_4 = 0.00$

Note that although the Law of Armed Conflict and Just War sources for the principles used in the REV model considered the *Prospect of success* principle to be important, its weight in the calibrated REV model is 0.00. This means that in order to best match the experts, the REV model treated violations of the *Prospect of success* principle as always less important *relative to the other ethical principles* considered in the model.

The MOE model was recalibrated to the rater data collected in this experiment. Details of the MOE model's calibration are omitted (since the model's full formulation is not given here), but a brief overview is provided. In terms of tangible consequences, civilian casualties accounted for 51% of the MOE model's valuation of a COA. This was followed by friendly utilities facilities destroyed at around 40% and all other consequences accounting for the remainder. Also of note is that this recalibrated MOE model regarded an actor's intention as much less important than the previous version did.

6. Results

This section reports the results of the calibration process, including discussions of the calibrated REV and MOE models, as well as how these models and all other rater classes fared when compared to the experts.

Table 3 shows the results of the iteration of the calibration procedure with the best performing set of REV model principle weights. As described earlier, this iteration of the procedure included 1000 partitions of the expert raters into standard and contestant raters and a comparison of the raters in each of the contestant classes to the standard raters for each partition. Each row of the table corresponds to one class of raters; the six base classes and the two composite classes of raters are included in the table. The first column identifies the rater class and the second column reports the number of raters in the class. The third column reports the total number of observations made for raters of that class. The fourth column contains the $\bar{\Gamma}_R$ values, which are the sample means of the $\Gamma_{R,i}$ values for all the observations involving raters in the class; the $\bar{\Gamma}_R$

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

Table 3. Results from the calibration process, with classes of raters ordered by descending $\bar{\Gamma}_R$ (mean agreement of the class with all experts).

Class of raters	Raters in class	Observations	$\bar{\Gamma}_R$	s	P-value w.r.t. next base class
REV calibrated	1	1000	−0.0650	0.0515	< 0.001
MOE recalibrated	1	1000	−0.2442	0.0381	0.1699
Military experts	51	25,525	−0.2455	0.1012	< 0.001
All experts	86	41,000	−0.2515	0.1253	n.a.
All humans	141	98,000	−0.2557	0.1332	n.a.
Non-experts	55	55,000	−0.2589	0.1390	0.1414
Humanities experts	35	17,475	−0.2603	0.1534	< 0.001
Random	1000	1,000,000	−0.2727	0.1035	< 0.001
REV dominant p_1	1	1000	−0.3783	0.0409	< 0.001
REV all weights equal	1	1000	−0.6438	0.0485	n.a.

values quantify the overall agreement of a class of raters with the class of expert raters over 1000 partitions. The fifth column reports the sample standard deviation for the $\bar{\Gamma}_R$ values for the class. Finally, the last column reports the P-value for the statistical hypothesis test comparing each class's agreement with that of the class below it. Note that three rows in the table do not have P-values. The all experts class and the all humans classes were not included in the class-to-class hypothesis test comparisons because they are composite classes that include raters from the base classes, and thus do not constitute independent samples as required by the hypothesis test. The P-value reported for military experts is for comparing that class to non-experts, the next base class in the table. The last row in the table, for the REV all weights equal class, has no P-value because there is no class below it to compare to.

Recalling that larger values of $\bar{\Gamma}_R$ indicate better agreement, the rater classes in the table are listed from top to bottom in descending order of performance. The table shows that the calibrated REV model outperformed all other classes of raters in the degree with which its assessments agreed with all expert raters, including all expert raters themselves. The recalibrated MOE model's performance is next but is substantially worse than the calibrated REV model. Its $\bar{\Gamma}_R$ value is numerically slightly better than all experts, but the difference is not statistically significant (the P-value for comparing the two classes agreement values for equality is > 0.05). All experts, in turn, outperform non-experts, and non-experts outperform random raters. For $\alpha = 0.05$, the P-values in Table 3 indicate that the differences in agreement between each class and the next one was statistically significant in all but two cases, MOE recalibrated compared to military experts and non-experts compared to humanities experts.

As can be seen in Table 3, all of the rater classes' $\bar{\Gamma}_R$ values are less than 0, a value that indicates more disagreement (discordant pairs) than agreement (concordant pairs). However, even the all experts class itself has a $\bar{\Gamma}_R$ value less than 0, which is evidence of the degree to which the experts disagreed among themselves. With so much disagreement among all experts, the rater class with whom the other classes are compared, it is numerically difficult for any other rater class to achieve a $\bar{\Gamma}_R$ value above 0. The calibrated REV model has the best $\bar{\Gamma}_R$ value, that value is very close to 0, and it outperforms the next best rater class by a statistically significant margin, suggesting that the calibrated REV model does a usefully good job of modeling the experts' ethical preferences.

7. Findings and future work

Overall, a model that replicates a principles-based assessment by human experts of the relative ethical violation of two military courses of action was successfully developed and calibrated. The REV model, although very simple mathematically (a linear weighted sum of inputs) turned out to be rather accurate, its effectiveness deriving from the proper choices of principles and weights. This section discusses the findings of this research and potential future extensions of it.

7.1.1 Findings of this research. The primary finding of this research is that a quantitative model is able to replicate ethical tradeoffs made by military and humanities experts, agreeing with expert raters more than any other class of raters. This finding is similar to and congruent with the primary finding of the research behind the earlier MOE model: that the overall concept is viable and practical.[15]

The secondary findings of this research contrast with those of the earlier MOE work. Most importantly, this work indicates that harm to civilians is not an overwhelmingly dominant factor in realistic military scenarios, which are laden with context and may require deciding between two potentially controversial actions. In addition, human raters expressed much less consensus in this experiment than in the original MOE research. However, this

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

may be a consequence of the nature of the carefully constructed ethical dilemmas that the human raters faced in the calibration scenarios. In the end, the sheer difficulty in evaluating these dilemmas was what extracted clear tendencies on how raters make ethical tradeoffs when a dilemma presses them—when ''push comes to shove.''

Despite the apparent lack of consensus on the part of human raters, patterns were discernible in their assessments, suggesting trends behind how raters weighed principles in their minds as they assessed the ethical dilemmas presented to them. For these raters, while civilian nonmaleficence is an important factor, military necessity and maintaining only proportional harm also contributed to various degrees to raters' assessments. The REV model's calibrated weights arguably represent an abstract encoding of these considerations in the minds of the human raters.

The REV model itself is limited to simulating the decisions of human raters (by way of making similar assessments), and no claim is made that it models the human raters' actual thought processes in any way. However, rater comments did provide some insight into those thought processes. Many raters, regardless of their expertise, explicitly stated that they make decisions by evaluating principles or making other difficult tradeoffs. A humanitarian perspective is not limited to humanities experts, nor are military experts the only people to incorporate military goals in their thinking. Viewing difficult situations from multiple perspectives provides a more holistic picture. This ultimately leads raters to assess a situation using the whole of their background and knowledge, then to make a choice, which itself hints at the ethical tradeoffs they make. One rater stated that he/she ''struggled with balancing ethics with experiences/opinions,'' which, according to the rater, caused ''a general central tendency'' in the answers provided. Rater comments reinforce the idea that they make tradeoffs between ''morals, values, ethics, and principles,'' allowed ''values [to] dictate [their] response[s],'' and vying for actions that would promote ''a just war.''

7.1.2 Future work. The concept that underlies the REV model is clearly viable. The earlier MOE model, first conceived as thought experiment, was enhanced through further research, and that further research guided the development of the REV model. This research toward a better understanding of ethical decision modeling can be continued, whether by refining the REV model itself or by using that underlying concept to develop another ethical decision support model.

One rater comment indicated that he/she could ''often'' identify ''much better options (i.e., COAs)'' than those presented with the scenario. An enhanced set of scenarios, perhaps with more than two potential COAs—or even with COAs that could be constructed from the ground up by the rater—could provide a deeper exploration of the ethical trade space. Interviews with subject matter experts or veterans could help identify additional viable COAs. In addition, the ethical trade space could itself be expanded. The set of ethical principles used in this research was chosen from those referenced most often in a few key sources. Further research could incorporate more principles that are relevant to military decision analysis but were not used in this experiment. The REV could be tested in a different domain, using a principle set and scenarios appropriate for that domain. Some audiences have specifically suggested that the concept could be useful for law and medicine—two fields that, like the military, face pressing ethical concerns with significant effects.

As noted earlier, the *Prospect of success* principle (p_4) seemed to be relatively irrelevant to the experts surveyed, despite its inclusion in the sources from which the principles were drawn. Explaining this, and perhaps replacing *Prospect of success* in the REV model with another principle that would have a non-zero weight, is a matter of future work.

Moreover, the current formulation of the REV model assumes that the weighting, or importance, of each principle is independent of the situation at hand. However, from a philosophical standpoint, this assumption may be questioned; for example, self-defense may be more important than justice in one situation but not in another. Scenario-specific weights may improve the performance of the model. Exploring this issue is possibly one of the most important philosophically, in that it gets to the core of what the model is about, but defining a comprehensive set of scenario classes and defining principle weights for each class is likely to be very difficult.

The technical effectiveness of the model is not the only concern. Because the model addresses ethical issues, it has the potential to be misused, e.g., by manipulating the weights to justify a questionable COA. The authors are well aware of this danger, and have addressed it from a philosophical perspective in related work.[14] More research needs to be done in this area in order to solidify the proper scope of the model.

Acknowledgments

Gregory B Tackett provided the key idea for the original ''MOE'' concept, which served as the precursor to and inspiration for this research. We are also grateful for the many participants who donated their time, assessments, perspectives, and expertise by responding to the surveys. Finally, the comments of the anonymous referees on an earlier version of this article led to substantial improvements to it.

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Milgram S. Behavioral study of obedience. *J Abnormal Social Psychol* 1963; 67: 371–378.
2. Zimbardo PG. A situationist perspective on the psychology of evil: understanding how good people are transformed into perpetrators. In: Miller A (ed) *The social psychology of good and evil: understanding our capacity for kindness and cruelty*. New York: Guilford, 2004, pp.21–50.
3. Baron-Cohen S. *The science of evil: on empathy and the origins of cruelty*. New York: Basic Books, 2011.
4. Baron-Cohen S and Wheelwright S. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J Autism Dev Disorders* 2004; 34: 163–175.
5. Simon RI. Should forensic psychiatrists testify about evil? *J Am Acad Psychiatry Law* 2003; 31: 413–416.
6. Knoll JL. The recurrence of an illusion: the concept of 'evil' in forensic psychiatry. *J Am Acad Psychiatry Law* 2008; 36: 105–116.
7. Welner M. Classifying crimes by severity: from aggravators to depravity. In: Douglas J, Ressler R and Burgess A (eds) *A crime classification manual*. San Francisco, CA: Jossey-Bass, 2007, pp.55–72.
8. Stone MH. *The anatomy of evil*. Amherst, NY: Prometheus Books, 2009.
9. Irwin T. *Nicomachean ethics*. Indianapolis, IN: Hackett Publishing, 2000.
10. Schelling TC. *The strategy of conflict*. Cambridge, MA: Harvard University Press, 1980.
11. Gips J. Towards the ethical robot. In: Ford KM, Glymour C and Hayes PJ (eds) *Android epistemology*. Cambridge, MA: MIT Press, 1995, pp.243–252.
12. McLaren B. Computational models of ethical reasoning: challenges, initial steps, and future directions. *IEEE Intell Syst* 2006; 29(4): 29–37.
13. Anderson M, Anderson S and Armen C. Towards machine ethics: implementing two action-based ethical theories. In: *Machine ethics*, Menlo Park, CA: AAAI Press, 2005, pp.1–16.
14. Reed GS and Jones N. Toward modeling and automating ethical decision-making: design, implementation, limitations, and responsibilities. *Topoi* 2013; 32: 237–250.
15. Reed GS, Tackett GB, Petty MD, et al. A model of 'evil' for course of action analysis. *Mil Oper Res* 2013; 18(4): 61–76.
16. Dawes RM. The robust beauty of improper linear models in decision making. *Am Psychologist* 1979; 34: 571–582.
17. Dawes RM. A case study of graduate admissions: application of three principles of human decision making. *Am Psychologist* 1971; 26: 180–188.
18. Dawes RM and Corrigan B. Linear models in decision making. *Psychol Bull* 1974; 81: 95–106.
19. Goodwin P and Wright G. *Decision analysis for management judgment*. Chichester, UK: John Wiley and Sons Ltd, 2004.
20. Ross WD. *The right and the good*. Oxford: Oxford University Press, 1930.
21. United States Air Force. Professional development guide. AF Pamphlet 36–2241, United States Air Force, July 2007.
22. Mill JS. Utilitarianism. 7th ed. In: *Fraser's magazine*. London: Longmans, Green, and Co, 1879.
23. Kant I. *Fundamental principles of the metaphysic of morals*. translated by Abbot TK. Charleston, SC: BiblioBazaar, LLC, 2011.
24. Beauchamp T and Childress JF. *Principles of biomedical ethics*. Oxford: Oxford University Press, 2001.
25. Gillon R. Medical ethics: four principles plus attention to scope. *BMJ* 1994; 309: 184.
26. Gillon R. Ethics needs principles—four can encompass the rest—and respect for autonomy should be 'first among equals'. *J Med Ethics* 2003; 29: 307–312.
27. DiMeglio RP, Condron SM, Bishop OB, et al. In: Johnson WJ and Gillman AD (eds) *Means and Methods of Warfare*, http://www.loc.gov/rr/frd/Military_Law/pdf/LOAC-Deskbook-2012.pdf (2012, accessed 2013).
28. Canadian Office of the Judge Advocate General. *Law of armed conflict at the operational and tactical level*, http://www.forces.gc.ca/jag/publications/oplaw-loiop/loac-ddca-2004/index-eng.asp (1999, accessed 2013).
29. *Catechism of the Catholic Church*. Vatican City: Libreria Editrice Vaticana, 2000.
30. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932; 22(140): 5–55.
31. Goodman LA and Kruskal WH. Measures of association for cross classifications. *J Am Stat Assoc* 1954; 49: 732–764.
32. Popping R. On agreement indices for nominal data. In: Saris WE and Gallhofer IN (eds) *Sociometric research. Volume 1. Data collection and scaling*. New York: St. Martin's Press, 1998, pp.90–105.
33. Brase CH and Brase CP. *Understandable statistics: concepts and methods*. Boston MA: Houghton Mifflin, 2009.

Author biographies

Gregory S Reed is a Senior Research Scientist at the University of Alabama in Huntsville's Center for Modeling, Simulation, and Analysis. In 2013, he completed a PhD in Modeling and Simulation, the first student to do so at UAH. Dr Reed's primary areas of research involve connecting technology with the human element, such as in data visualization, cognitive and behavior modeling, and training.

Mikel D Petty is Director of the University of Alabama in Huntsville's Center for Modeling, Simulation, and Analysis and an Associate Professor of Computer Science. He received a PhD in Computer Science from the University of Central Florida in 1997. Dr Petty has worked in modeling and simulation research and education since 1990 in areas that include verification and validation methods, simulation interoperability and

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016

composability, human behavior modeling, and simulation software architectures. He has published over 185 research papers and has been awarded over $16 million in research funding. He served on a National Research Council committee on modeling and simulation, is a Certified Modeling and Simulation Professional, and is an associate editor of the journal *Simulation*. He has served as dissertation advisor to five graduated PhD students, including the first two students to receive PhDs in Modeling and Simulation at Old Dominion University and the first student to receive a PhD in Modeling and Simulation at UAH.

Nicholaos J Jones earned his PhD from the Ohio State University and currently is an Associate Professor of Philosophy at the University of Alabama in Huntsville. His research focuses on the role of idealizations and diagrammatic representations in the natural sciences. He has published in a variety of journals, including *Studies in History and Philosophy of Modern Physics*, *Biology and Philosophy*, *Journal for the General Philosophy of Science*, and *Erkenntnis*.

Anthony W Morris received his PhD in Experimental Psychology from the University of Connecticut, specializing in Human Factors, Ergonomics and Performance Learning. He is currently Chief (A) of the Aviation Missile Command Field Element in the Army's Research Laboratory Human Research Engineering Directorate. His research focuses on pilot performance modeling and assessment for utility rotorcraft operations. His expertise covers issues of information detection and action control for navigation and instrumentation operation. He also has extensive experience with virtual and constructive simulations involving digital human models for life cycle management for manufacturing, maintenance, equipment operations as well as test and evaluation.

John P Ballenger is Associate Research Professor of Management Science and Principal Research Engineer at Research Institute, University of Alabama in Huntsville (UAH). Dr Ballenger came to UAH in 2007 after spending the past 27 years with Raytheon Missile Systems. Dr Ballenger retired from Raytheon as a Senior Program Manager. Dr. Ballenger won several awards at Raytheon, including the company's inventor's award in 1986. He was the Huntsville Association of Technical Societies' professional of the year in 1998. Dr. Ballenger received the Outstanding Undergraduate Teaching Award from UAH for 2009–2010. In addition to his teaching experience at UAH, he has been adjunct faculty at Embry-Riddle Aeronautical University, Georgia State University and Boston University. He received his PhD and MEd from Georgia State University and a BIE from Auburn University. Dr. Ballenger is a distinguished military graduate from Army ROTC, and he was commissioned in the Regular Army, Field Artillery, in June 1969. He advanced to the rank of Captain in 1971, and he was released from active duty in June 1974 with an honorable discharge.

Harry S Delugach holds degrees from Carleton College, the University of Tennessee and his PhD from the University of Virginia. In addition to directing the Honors College, he teaches and supervises courses in software engineering. He is one of the pioneers in the development of conceptual graphs, a knowledge modeling formalism that has become one of the standards in ISO/IEC 24707 (2007) Common Logic, for which he served as editor. His research interests are in intelligent decision-support and knowledge acquisition systems, as well as software requirements development.

Downloaded from dms.sagepub.com at UNIV ALABAMA HUNTSVILLE LIB on March 28, 2016