A Model of ''Evil'' for Course of Action Analysis

Gregory S. Reed

University of Alabama in Huntsville, reedgs@uah.edu

Greg B. Tackett

US Army Aviation Test and Evaluation Command, Ballistic Missile Defense Systems Operational Test Agency and Ballistic Missile Defense Evaluation Directorate, gregory.b.tackett.civ@mail.mil

Mikel D. Petty and John P. Ballenger

University of Alabama in Huntsville, pettym@uah.edu, ballenj@uah.edu

APPLICATION AREAS:
Modeling, Simulation, and Wargaming
OR METHODS:
Decision Analysis, Simulation

ABSTRACT

Military planners must consider the undesirable secondary effects of military operations, such as civilian casualties, physical infrastructure damage, and societal disruption. A quantitative model that can be used to evaluate and compare the intentional harm, or ''evil,'' caused by alternative courses of action (COAs) would be useful to military planners. Two versions of a ''Metric of 'Evil,''' a model of the harm associated with military COAs intended to allow the comparison of COAs on an ethical basis, have been developed. The models consider both the results of a COA and the intentions of those executing it. The models were experimentally validated by comparing their assessments with those of human experts with backgrounds in ethics, religion, political science, and military history. Pairwise comparisons of the relative evil of pairs of COAs from a set of selected historical events were made by four sets of raters: human experts, human nonexperts, the models, and random raters. Accuracy in these assessments was defined as cumulative agreement with the experts. One of the versions of the model agreed with the human experts over all situations as well as all human raters. The results suggest that a quantitative model can capture the criteria and sensibilities used by human experts in assessing the evil associated with military courses of action.

INTRODUCTION

Regarding the evaluation of military courses of action (COAs), Tackett (2009) states that, ''evil is subjective … yet, real life demands humans regularly make choices based on the relative goodness … choosing the lesser of two evils.'' The development, analysis, comparison, and approval of COAs is an inherent part of the military decision-making process (FM 101-5, 1997). This paper proposes a model, informally coined the ''Metric of 'Evil,''' that could provide the staff and commander the ability to choose the lesser evil.

To develop the model, we first needed a measurable definition of evil. An example dictionary definition states that evil is, ''morally bad … wicked … harmful or tending to harm'' (Illustrated Oxford Dictionary, 2003). Zimbardo (2004) defines evil as ''intentionally behaving, or causing others to act, in ways that demean, dehumanize, harm, destroy, or kill innocent people.'' According to Staub (2004), ''[e]vil is an extreme and sometimes repeated form of people harming others.'' Baumeister (1999) defines evil as ''actions that intentionally harm other people.'' Two terms in defining evil became apparent: ''intention'' and ''harm.'' Thus, for an operational definition of evil to use in our experiment, evil was defined as ''intentional or anticipatable harm resulting from a given action.'' In our experiment, subjects, in general, weighted the harm to noncombatants more evil than harm to combatants. This finding reflects Zimbardo's (2004, 2007) definitions. Still, several subjects in our experiment did not agree with our definition of evil—satisfactorily defining evil is a difficult task. Miller (2004) states that ''the terms 'good' and 'evil' … are value-laden, perhaps even grandiose … have religious overtones, among many others.'' This study addressed evil in a military context, and in a given war or military conflict, the military personnel on all sides involved are charged with intentionally inflicting harm on their opponents. Baumeister (1999) argues that is hard ''to find a definition of evil that will satisfy everyone.''

PROBLEM STATEMENT AND APPROACH

A set of historical events was chosen, human subjects were obtained, and a design for the experiment was developed. Fact sheets were developed for all of the historical events selected. Then, as the experiment was conducted, Tackett's (2009) initial model was extended over several iterations. At the conclusion of the experiment, an analysis of results commenced. Based on this analysis, adjustments to the model were made to improve the agreement between the model and the human experts.

It is envisioned that, once calibrated sufficiently, the model may be used by military staff members in conjunction with other decision criteria in the comparison of COAs. In the future, ''evil'' or ''intentional harm'' might be included among decision criteria. The UAHuntsville model software implementation, although not a field-ready software product at present, has design features intended to support its conversion to a robust tool that could be used in the field.

LITERATURE REVIEW

This project involves several disparate subjects—soliciting experts' opinions, quantifying harm, and aiding in COA analysis, for example. There is abundant literature on most of these subjects when they are each taken in isolation; however, there is little precedent for the specific topic under study. Thus, this review concentrates on the research within these subjects that has the most significant impact on this project.

Egan (2007) defined a COA as ''a possible plan open to an individual or commander that would accomplish, or is related to the accomplishment of the mission.'' Essentially, a COA is a series of actions that may potentially result in a set of desired outcomes (Egan, 2007). COA and scoring techniques involve evaluating COAs to determine which are most effective; that is, to determine those that lead to optimal outcomes based upon the commander's intent. Effects-based operations practices aim to evaluate secondary, tertiary, and other indirect effects of a COA. These effects may not necessarily arise within a particular military conflict; rather, some may arise well within the postconflict phase (Batschelet, 2002). Many sources illuminate just how important it is to keep these far-reaching effects in mind when evaluating COAs (Egan, 2007; Hanna, 2004). Egan (2007) illuminates just how modeling and simulation (M&S) techniques enhance this level of analysis, stating that, ''Metrics must be established that adequately describe and quantify the relative merits of such disparate COAs.'' The goal of COA analysis, then, is to determine the set of actions, forceful or not, that best lead to a set of desired outcomes and avoid negative effects (Batschelet, 2002). The purpose of our particular model is to provide an ethical viewpoint from which COAs can be evaluated. Although the set of ethical implications of a COA is only one of many factors to consider, the model's results are intended to be factored into a commander's decision.

Decision matrices are one tool that allows decision analysts to frame their set of alternatives. This is important because Heuer (1999) states that, when provided with a set of information items, experts are largely unaware of which items of information dominates their own conclusions. The procedure of developing and populating a decision matrix has seen application in COA analysis with mixed results (Fallesen, 1995). Although decision matrices can be somewhat limited, the model's formulation is inspired by their structure.

The end goal of a decision model is often to provide the necessary level of understanding to decide on a given COA. In contrast, Goodwin (2004) describes L.D. Phillips' concept of the requisite decision model, the goal of which is not to provide decision makers with a final decision, but to allow them to resolve potential inconsistencies in their thought processes. We expect that analysts who use the model, when used to evaluate hypothetical future courses of action for a given situation, will attempt to relate their situation to comparable historical situations—which is most effective when their experience provides them with a number of potential analogues and when they can determine the similarities and differences between the situation at hand and any comparable situations (Heuer, 1999).

The model accepts as input numerical details of the consequences of the COAs, performs computation on those details, and produces an absolute numerical measure of the associated evil. Perhaps because of these difficulties, the literature related to quantifying evil is small, and not always directly relevant. Welner (2006), for example, provides a lengthy discussion of how and why crimes are classified by severity. Moreover, evil may involve more than harm to human beings. Damage to a society's infrastructure harms that society and the people within it, and arguably should be included in a quantification of the evil associated with a military COA. Richardson (2004) proposes a model of postconflict reconstruction to numerically capture several varied infrastructure factors.

The process of validating the model through assessment of historical events was crucial to the development of our model. A significant portion of this process relies upon reliably soliciting expert assessments of these historical events. Therefore, we also researched techniques for effective presentation of questions to experts, distribution of questionnaires, and possible psychological biases that may affect the results of our validation process. Babbie (2002) presents several suggestions. First, he recommends a pretest as a way to ensure that one's survey methodology solicits appropriate responses from participants clearly. Second, he recommends randomizing the order of questions to help alleviate question order bias. Third, allowing surveys to be filled out and returned by the easiest means possible will ensure that more participants provide responses. Fourth, he recommends short, concise questions. These suggestions helped shape the survey provided to experts.

In discussing analysis biases in general, Heuer (1999) states that making such biases explicit, and interpreting results with those biases in mind, leads to a more objective analysis than does an attempt to simply cover or suppress them. A complete avoidance of psychological biases is virtually impossible, and these biases have the potential to disrupt a survey. For example, social desirability may bias participants' responses (Babbie, 2002).

MODEL DESIGN

Following is a presentation of both the initial design of the model and an extended design. The extended model, the ''Metric of 'Evil''' v1 (or ''Metric v1''), was heavily influenced by the initial version of the model, the ''Metric of 'Evil''' v0 (or ''Metric v0''). Thus, the models have several elements in common. Both are designed to compare the evil associated with COAs via quantifiable input parameters that represent facets of each COA that have an ethical impact, including death and hardship. The models utilize a similar definition of evil at their core, which is based upon manifested evil intention. In addition, both models account for the intent of the actor as well as the results of his or her actions—a death that is deemed necessary or accidental, for example, is viewed as less evil than one that is malicious or absolutely intentional.

In both versions of the model, the calculation of a COA's evil is a sum total of the evil associated with each of its tangible ethical aspects. Although both versions of the model necessarily incorporate value judgments, the decisions that an analyst must make when providing input to Metric v1 concentrate on more observable phenomenon. Metric v1 intends to further separate the subjectivity involved in value judgments from the user's input; the premise is that the user may then focus on providing more objective information so that reproducible comparisons can be more readily made. Many of the subjective aspects of Metric v1 may be modified to incorporate vastly different value systems and philosophical points of view; however, analysts need not take these aspects into account when using the tool.

Metric v0 Design and Implementation

This model measures the harm that results from a given action and does not attempt to capture evil thoughts alone. Essentially, it calculates the evil associated with a particular COA by the degree of harm done to individuals, weighted by the level of malicious intent of the actor. These factors are captured by categorizing harm based upon two criteria: the *order of evil* and the *harm index*. The Orders of Evil include:

1. *Necessary evil*: Acts that are considered to be morally necessary, such as those that result from self-protection or the protection of innocent individuals. Examples include imprisonment of criminals and most hardship to soldiers in a military context.
2. *Consequential evil*: Acts that the actor may regret, but that result from necessary actions. These include collateral damage and incidental civilian casualties.
3. *Selfish evil*: Acts that constitute preventable harm resulting from inaction.
4. *Malicious evil*: Evil acts that are considered to be unnecessary and morally unjustifiable, including terrorism, genocide, and human rights violations.

The harm indices include hardship, suffering, injury, and death. Hardship is differentiated from suffering in that the latter is more severe. Examples of hardship may include those captured as prisoners of war, where suffering may include those left without essential resources. Injury and death are straightforward.

The model is characterized by an arrangement of these two factors on separate axes in a matrix format. The design uses four degrees of severity across both axes, resulting in 16 entries in a Matrix of Evil. The number of individuals inflicted with each degree of harm and with each specific degree of intent is placed into a Matrix of Evil in the appropriate entry x_{ij}. The Matrix of Evil for a COA that results in 2,000 malicious deaths and 3,000 necessary injuries, for example, will have $x_{44} = 2{,}000$ and $x_{13} = 3{,}000$.

The evil associated with a COA is then calculated as a weighted sum of each entry into the matrix. The weights used for each entry, known as the scale factors, represent the overall severity of the harm and intent associated with the entry. The scale factor for each entry x_{ij} is calculated as H_j^{Ei}—that is, the harm index raised to the power of the order of evil. In this formulation, malicious harm is considered to be much more severe than necessary harm since there is an exponential relationship between evil intent and overall evil. Thus, for example, in calculating the overall evil associated with a COA—its *sum of all evil index*—the number of necessary sufferings is multiplied by 2, and the number of malicious deaths is multiplied by 256 and, therefore, one malicious death is deemed equivalent to 128 necessary sufferings. Overall, this sum of all evil index is calculated as $E_s = \mathbf{\Sigma}\, H_j^{Ei}\, x_{ij}$.

The *Delta goodness* between COAs represents a comparison of the evil associated with two COAs. The Delta goodness between COAs *A* and *B* is calculated as $E_s(A) - E_s(B)$.

Metric v1 Design and Implementation

The next design iteration, Metric v1, shares some design aspects with Metric v0 but differs with it in others. This section describes the design and implementation of Metric v1, including an overview of the design, input variables and their weightings, global control parameters for the model, and the series of calculations that the model uses to compare COAs.

Design overview. Like Metric v0, Metric v1 captures evil in terms of intent and harm. Whereas Metric v0 contains a ''harm index'' and ''order of evil'' to frame these, Metric v1 captures harm via magnitudes of particular countable tangible consequences and intent through other model factors.

Metric v1 determines the relative evil associated with a particular COA by soliciting input parameters from an end user, incorporating certain modifiable global factors. This rating can then be used to compare the ethical implications of two COAs under review, providing another formulation for a comparative Delta goodness between COAs.

Because of the subjective nature of morality in general, the model is designed to be flexible with respect to ethical factors; that is, it can be modified to reflect different baseline morality systems. This is an important facet to capture due to cultural differences in the perception of morality—for example, one culture may deem intention to be especially important, where another culture may concentrate solely on a utilitarian, results-based view of an action's end results; one culture, which may view mosques and other cultural sites as sacred, may prioritize protecting them over protecting the lives of soldiers and civilians, where another culture may prioritize the protection of life over all other factors. The default values for the model's various weights and factors attempt to represent the assessment of individuals who are largely American in culture and who are experts in military, ethics, philosophy, psychology, history, and other related fields.

The input variables represent the potential results—specific types of harm and intent—of a COA. From these inputs, a measure of evil associated with the COA under study is calculated.

Given that the results of a COA are inexact estimates, values for the input parameters take the form of an estimated low value and high value. These values represents the range that the input parameter's end value is expected to take. The Delta goodness is adjusted based upon the user's confidence in his or her estimations.

However, evil may not necessarily scale linearly with the results of a COA. Conceptually, this is similar to the economic law of diminishing returns—one additional dollar is worth less, in a psychological sense, to a millionaire than to an impoverished individual. Similarly, the marginal evil for each additional life lost or person enduring hardship may be said to decrease as more individuals are affected; effectively, a COA in which two lives are lost may not be twice as evil as a COA in which one life is lost in some cultural contexts. Thus, the model accounts for the possibility that, psychologically, a given culture's baseline morality in terms of its perception of evil may not scale linearly with unethical results. Once a COA begins to lead to loss of life, it may be seen as more evil than one that does not; to this particular culture, the exact number of lives lost may be less significant as the number increases.

The intention associated with a COA's level of harm is also provided, but in a different sense than that described in the original model. Unexpected results are rated as very low intention, anticipated results are rated as a middling intention, and results that the COA is intended to cause are rated as a high intention. An actor may determine that he or she intends to cause enemy combatant casualties in evaluating a particular COA; in another, he or she may see enemy combatant casualties as a factor that is anticipated but not directly intended. Unexpected results are considered by the model only when evaluating historical events—all future results, by definition, are considered to be at least anticipated.

Input variables and their attributes. Table 1 provides the input parameters used by the model. These input parameters were influenced by those in the original formulation of the model and further refined by the researchers' own expertise with respect to military operations and through the literature review process. The following should be noted with respect to the chosen input parameters. ''Friendly'' and ''Enemy'' are

Table 1. Metric v1 input parameters.

Category	Parameter	Unit
Friendly force casualties	Killed	Persons
	Wounded or injured	Persons
	Captured or missing	Persons
Enemy force casualties	Killed	Persons
	Wounded or injured	Persons
	Captured or missing	Persons
Noncombatant casualties	Killed	Persons
	Wounded or injured	Persons
	Captured or missing	Persons
Noncombatant hardship	Left without essential facilities/resources	Persons
	Homeless or refugee	Persons
	Unemployed	Persons
	Economically damaged	Persons
Friendly infrastructure damage	Essential facilities destroyed	Count
	Cultural facilities destroyed	Count
	Nonessential facilities destroyed	Count
Enemy infrastructure damage	Essential facilities destroyed	Count
	Cultural facilities destroyed	Count
	Nonessential facilities destroyed	Count
Neutral infrastructure damage	Essential facilities destroyed	Count
	Cultural facilities destroyed	Count
	Nonessential facilities destroyed	Count
Moral/ethical/legal considerations	Major international law violations	Count
	Major treaty violations	Count
	Minor international law violations	Count
	Minor treaty violations	Count
	National promises broken	Count

defined with respect to the user of the tool. ‘‘Force’’ casualties represent casualties of combatants, soldiers, and other fighters; noncombatants represent civilians and other nonfighters. Essential facilities include those that provide necessary resources to a population, including hospitals and power stations. Cultural facilities are facilities that impact a national or group culture. These include churches, monuments, and historical sites. Nonessential facilities include other infrastructure, such as sports stadiums and businesses, not listed as essential or cultural facilities. Major treaty and international law violations represent those that have a significant chance of leading to retaliation, civil war, or other harmful actions, where minor violations do not. Some data estimates may be applicable to more than one category; for example, a particular group of noncombatants may be both captured and left unemployed. In this case, the estimate should be placed in the category nearest the top of the list; the model's weights have been determined with this assumption in mind.

The entire collection of input parameters is intended to capture direct harm to individuals (e.g., casualties and hardship) as well as indirect factors that have the potential to cause significant future harm (e.g., treaty violations that may result in future conflicts in which lives are lost). Cultural facilities, such as churches, are included and differentiated from nonessential facilities due to the significant impact that their destruction may have in some morality systems.

The following attributes are associated with each input parameter i:

- l_i and h_i: The low and high estimated values, respectively, constitute the lowest and highest reasonable values that the input parameter will have.
- c_i: The confidence level is a measure of the user's confidence that the actual value for the parameter will reside between l_i and h_i. A value of 1.0 represents complete and full confidence in the values estimated above. A confidence of 0.9 may, for example, reflect a value derived from primary source materials, where a confidence of 0.5 may reflect a value that is calculated or estimated from a set of relatively uncertain data.
- m_i: The measure of intentionality is a measure of how directly the COA under study causes this particular result—the intention of the actor. Actions that are directly caused and intended to occur are considered to be completely direct and intended, where indirect effects may include enemy retaliation and other similar, less foreseeable actions. The model solicits intentionality from the user as one of three potential values and maps them to a value used in its internal calculations. Unexpected results, intentionality of 0.1, refer to cases where the actor under study has no reasonable way to assume that its actions would cause the result associated with this parameter. This category may only be used when evaluating historical events, as only anticipated or intended results can be recorded by the model for future potential COAs. Anticipated results have an intentionality value of 0.5, and represent those that are or can be foreseen by the actor, but are not necessarily caused by that actor. Intended results are those that are seen as a direct result of the COA under study—that is, results that the COA is essentially designed to achieve—and have an intentionality value of 1.0.
- w_i: The weight is the influence of this input parameter in deciding the overall evil associated with the COA, normalized such that they sum to 1.

Control parameters. Certain global factors influence how the model calculates its end result:

- F: The evil power factor determines the degree to which intention is factored into the evil associated with each parameter. For $0 \leq F < 1$, there is little difference between anticipatable harm and intentional harm. As F approaches zero, intention has a less significant impact on the overall evil rating. For $1 < F < \infty$, there is a large difference between anticipatable harm and intentional harm. As F approaches infinity, only those results that are fully intended by the actor will be factored into the assessment. $F = 1$ represents a linear relationship between intention and evil. For very low values of F, an actor's intention is barely factored into the model's analysis—the evil associated with an action is nearly constant with respect to how intentional its results are. In this case, anticipated evil results are considered to be nearly as evil as intentional results within the same category. For *F* values that are very high, the opposite is true; only the evil results that the actor absolutely intends will be considered evil, and unforeseen and anticipated results will be overshadowed by the intended results. Middling *F* values represent some mixture of these two philosophies, a certain degree to which the intention of an action matters in an ethical sense. The exact value of *F* used depends upon the baseline morality that the model represents.

- *D*: The diminishment factor captures the effect of the numerical magnitude of harms. In a certain baseline morality, one's perception of evil may not scale linearly with the sheer number of individuals or items affected by a COA. The diminishment factor captures this by scaling input parameter values appropriately. The effect of the diminishment factor on the input parameter values is similar to the effect of the evil power factor on intention. When *D* is very low, the marginal evil of additional quantities decreases—the fact that civilian casualties occur will outweigh the actual number of civilian casualties, for example. When $D = 1$, there is a linear relationship between an input parameter's value and the evil associated with it.
- Z_l and Z_h: The low and high confidence standard scores allow low and high estimates of input parameters to be reconciled into a single value. For a given COA, a range of evil is calculated. This range is based on the low and high estimates for input parameter values. An overall confidence in this range of evil is also calculated based upon the analyst's confidence in his or her input values. Together, this range of evil and the confidence in this range identify a particular statistical distribution (assumed to be a normal distribution) that the evil value may have. This accounts for the possibility that an analyst may not be completely confident in the estimates provided to the model and that there is a certain level of uncertainty in what the evil associated with the COA will be when it is actually carried out. If the overall confidence in the range is very high, then it is assumed that there is a significant chance that the actual evil of the COA resides within that range. If the overall confidence is low, then there is less of a chance that the COA's evil value resides within the range. The high confidence standard score represents half the number of standard deviations from the mean that the range covers if there is a 100 percent confidence in the range. The low confidence standard score represents the same for a 0 percent confidence in the range. For example, if the confidence in the range for COA *j* is 100 percent, and if $Z_h = 3.0$, then the standard score $z_j = Z_h = 3.0$. The range calculated for the COA thus represents $\pm 3\sigma$. This, in turn, means that there is a 99.7 percent chance that the evil value for the COA falls within the range calculated by the model. In calculations where confidence in the COA's evil value range is low, the range calculated by the model will represent less of the span of potential evil for the COA. The actual standard score used in calculations is linear, bounded by these two values, and dependent on the actual confidence calculated for the range. Rather than relying upon simple averages of low and high estimates for values, this formulation compares COAs with uncertainty in mind. All other factors being equal, the COA that has more of a potential for evil is thus determined by the model to be more evil overall than another.

Evil calculation formulas. Steps taken to calculate the potential evil associated with a particular COA *j*, reported as a mean and standard deviation, are as follows:

Enumeration seems to be lost here as well.

1. Gather low estimated values, high estimated values, and confidence values for each input parameter. Weights and global parameters are taken as given.
2. Calculate normalized weightings for the input parameters: $w_{ni} = w_i / \Sigma\, w_i$.
3. Calculate the low evil, high evil, and mean evil for the COA:

$$e_{lj} = \Sigma l_{ij}^{D} m_{ij}^{F} w_{ni}, e_{hj} = \Sigma h_{ij}^{D} m_{ij}^{F} w_{ni},\ \mu_j$$
$$= \left(e_{lj} + e_{hj}\right)/2$$

Table 2. Historical events and situations used in the validation experiment.

Event		Situations	
A	Warsaw Uprising of 1944	A1	Germans
		A2	Poles
		A3	Soviets
B	Korean Air Lines Flight 007	B1	Non-Soviets
		B2	Soviets
C	Bay of Pigs invasion	C1	Cuban communists
		C2	Cuban exiles
		C3	USA/CIA operatives
D	Operation Enduring Freedom	D1	US military
		D2	Taliban
		D3	Non-Taliban Afghanis

The range from low evil to high evil represents the range of potential evil for that COA. In this formulation, the following mathematical convention is used: if $m_{ij} = 0$ and $F = 0$, then $m_{ij}{}^F = 0$. This allows the evil power factor to function as expected when it is equal to zero.

4. Calculate the confidence for the range of evil: $c_j = \boldsymbol{\Sigma}\, c_{ij} w_i$.
5. Calculate the standard deviation for the evil. This is based upon the proportion of the distribution of potential evil that the range covers, which is in turn based upon this confidence. The standard score (or z-score) for the distribution can be calculated as follows: $z_j = Z_l + (Z_h - Z_l)c_j$. The typical calculation for the standard deviation of a normal distribution would be the following: $\sigma_j = (e_{hj} - \mu_j)/z_j$. However, in the case where $e_{lj} = e_{hj}$, which arises when the user is evaluating historical scenarios where exact effects are known, the standard deviation for the distribution would be zero. This would lead to a meaningless formulation for the Delta goodness. Therefore, the following calculation is used instead: $\sigma_j = \max(\varepsilon, (e_{hj} - \mu_j)/z_j)$. The variable ε is chosen to be very small—10^{-5} is the default quantity. This formulation for the standard deviation ensures that there is at least some meaningful deviation in the distribution, allowing two COAs to be compared.

The mean and standard deviation of potential evil associated with the COA are reported as μ_j and σ_j, respectively. The Delta goodness between two COAs is a measure of how much less evil one COA is than another. Thus, it can be taken as a measure of the statistical distance between two COAs' distributions of potential evil. Steps taken to calculate the relative evil associated with a particular pair of COAs j and k is as follows:

1. Calculate the mean and standard deviation for the potential evil associated with each COA—μ_j and σ_j, μ_k and σ_k.
2. Calculate the Delta goodness via the distance between the two distributions:

$$\Delta G_{jk} = \frac{\mu_k - \mu_j}{\sqrt{\sigma_k^2 + \sigma_j^2}}$$

This formulation assumes that the evil for COAs j and k were calculated using the same set of global factors and input parameter weights. COAs j and k cannot be compared, for example, if their calculations use different diminishment factors or if any input parameter weighting differs between them.

Extensions of Model v1 and alternative considerations. This model design has a few key differences from its predecessor. First, the weightings associated with specific types of harm (casualties, economic hardships, legal violations, etc.) and the degree to which intention affects the model's calculations can both be adjusted—for example, to reflect different value systems. Although these parameters are not intended to be modified by analysts who use the model, separating the weightings from the model design allows for a great degree of flexibility. Second, inputs into the model have been clearly

Table 3. Number of human raters who responded to each event pair.

Event pair	Survey order	Military experts	Nonmilitary experts	Nonexperts	Total
AB	**AB**	2	1	1	9
	BA	1	1	3	
AC	**AC**	2	1	1	8
	CA	1	2	1	
AD	**AD**	2	1	2	10
	DA	3	1	1	
BC	**BC**	1	1	1	7
	CB	2	1	1	
BD	**BD**	1	2	1	8
	DB	2	1	1	
CD	**CD**	2	1	1	8
	DC	1	2	1	
Total		20	15	15	50

defined and involve as few subjective value judgments on the part of the model's users as possible. Measuring an actor's intention allows the model to capture a baseline morality in its assessments while clarifying user input. A certain level of subjectivity in categorizing inputs into a model of this nature is unavoidable, but clearer definitions for the categorizations allow for a more objective analysis. Thus, Metrics v0 and v1 present differing approaches to balancing the handling of subjective factors with the need for an objective perspective of a COA's potential results.

VALIDATION EXPERIMENT

This section explains the design and conduct of the experiment used to validate, or more precisely, calibrate the models. The overall validation method, the process of selecting, studying, and documenting the historical events used for validation, the process used for gathering and analyzing the validation data provided by the human experts, and the details of calculating agreement between the humans and the implemented models are discussed in turn.

Introduction

For the proposed model to be useful, it must provide ratings of the evil associated with a COA that are, in some useful sense, accurate. The meaning of ''accurate'' in this context is both important, in that it drives the design and conduct of the validation experiment, and nonobvious, in that it is not easy to define. This leads to two questions with respect to validating (or calibrating) the model: first, what will the evil ratings produced by the model be compared to; and second, how will the comparisons be performed?

Regarding the first question, the assessments of human experts were chosen as the standard of comparison for two reasons. First, in operational use the model is intended to replicate and potentially stand in for the assessments of such persons, so comparing the model's ratings to the experts' assessments during validation (calibration) appeared to be the most straightforward approach. Second, no other viable standard for comparison seemed to be available; there is no physics equation, authoritative data source, previously existing model, or accepted body of precedent for such assessments. The experts were simply assumed to be correct, and the model validated (calibrated) based on how well it matched the experts.

Regarding the second question, the method was based on the general model validation method known as retroactive prediction, or ''retrodiction,'' wherein a model is used to simulate a historical event, and the model's results are compared to the historical outcome; this method is often used for combat models (Barbosa, 2010). Specific historical events were selected, studied, and documented. In particular, the data values

Table 4. Overall ranking and agreement results.

	Metric v1 weights, situation pairs under study				
	All situations		All situations except Soviets in KAL and Cuban communists in Bay of Pigs		
Rater class	Average Rank	Average Agreement	Average Rank	Average Agreement	Count
Military experts	30.9	0.5960	35.0	0.5766	20
Nonmilitary experts	24.7	0.6437	23.1	0.6584	15
All experts	28.2	0.6164	29.9	0.6116	35
Nonexperts	25.9	0.6340	22.1	0.6592	15
Random	57.7	0.3966	57.3	0.3870	12
Metric v1, skewed	**27.0**	0.6112	**31.0**	0.6373	1
Metric v1, unskewed	**37.0**	0.5618	**32.0**	0.6340	1
Metric v0, calibrated	**30.0**	0.6022	**39.0**	0.5816	1
Metric v0, original	**46.0**	0.4742	**44.0**	0.5065	1

needed as input to the implemented models for the selected events were extracted from the historical literature, a process that was unexpectedly difficult. Then the various classes of raters (human experts, human nonexperts, random raters, and models) were asked to assess the evil associated with the courses of action in those historical events, and the models' ratings were compared to those of the other classes.

Historical validation events

Seven historical events were considered for use in the validation experiment, from which a set for four was chosen. Those events, and the actors identified within each, are:

- Warsaw Uprising of 1944—German, Polish, Soviet
- Korean Air Lines Flight 007—Non-Soviet, Soviet
- Bay of Pigs invasion—Cuban communists, Cuban exiles, USA/CIA
- Operation Enduring Freedom—US military, Taliban, Non-Taliban Afghan

The set of selected historical events was considered to be broad enough to represent an adequate sampling of different types of military situations. Each event possesses characteristics that made its inclusion in the model appropriate and worthwhile. Moreover, the fact that multiple actors are present in each event provides a way to analyze a greater variety of actions.

The Warsaw Uprising of 1944 proved ideal for the validation. Of the four events selected, the uprising was the largest in terms of the sheer number of individuals involved. The sources provided information on combatant casualties, noncombatant casualties, and those noncombatants removed from their homes after the fighting had ceased. The event also possessed a passive actor, allowing evaluation of the evil of inaction. The immense damage done to Warsaw itself allowed data on infrastructure damage to be included.

The shoot down of Korean Air Lines (KAL) Flight 007 was not a classic military operation. Of the events chosen, this event produced the smallest number of casualties; therefore, in many respects, the background and aftermath were more important than the event itself. This presents a potential challenge to a model that relies more upon quantifiable data than upon context and thus provides a test of the model's flexibility.

The Bay of Pigs invasion was the most conventional military operation of the events selected, with an amphibious invasion followed by engagement of two forces. Detailed statistics were found for combatant casualties, whereas noncombatant casualties and infrastructure damage were relatively low. This event presents clear moral questions concerning the evil of invading a sovereign country and the effects of political considerations.

At the time of the study, Operation Enduring Freedom was a current event. Thus unlike the other events, its aftermath and context could not be known. Its ongoing nature tested the model's ability to analyze a modern military situation. Since the data associated with this event was more uncertain than the others, an analysis of it was like an analysis of potential COAs.

Each of the validation events was studied to acquire quantitative data for use in the model as well as a historical narrative containing relevant background, context, and accounts of the events themselves. Most quantifiable data used for the events was found directly within sourced material, but some was estimated based on numbers in other categories (e.g., the number wounded was estimated to be twice the number killed). Primary sources and scholarly books were the main source of information for the events. The large amount of information available for each event was reduced to a two-page summary for each event to be provided to the raters.

Experimental design

The objective of the experimental design was to generate data representing the opinions of the human experts and nonexperts the models could be compared to. The fact that a comparison was the objective was a driver of the experimental design.

Each of the four validation events involved two or three ''sides'' or ''factions'' competing or cooperating in the event. The combination of an event and a faction was termed a ''situation.'' For ease of record keeping, the events and situations were each assigned codes; the events and situations with their codes are listed in Table 2.

Both versions of the model return absolute numerical ratings of the evil associated with a COA. Given that human experts were to be used as the standard for comparison for validating (calibrating) the model, a naïve experimental design would have been to ask the human experts to produce numerical ratings in the same range (maximum and minimum values) as the models, and directly compare the numerical ratings of the experts to the models, perhaps using a simple statistical hypothesis test. However, after due consideration it was considered likely that the experts, working independently of each other, would use an absolute numerical rating scale differently and inconsistently from each other, even if they associated the same subjective degree of evil to a COA. (One need only consider the sometimes widely divergent scores assigned by judges in gymnastics competitions.) In contrast, asking human raters to compare the relative evil of two events was deemed much more reliable and consistent between two raters who actually held similar opinions. Consequently, the experimental design depends not on absolute ratings of evil associated with situations from the raters, but on relative ratings of the evil in a pair of situations. To that end, the human raters were each asked to perform a series of pairwise comparisons (i.e., to assess the relative evil of pairs of situations).

Given four events (A, B, C, and D), there are six possible pairs of events (AB, AC, AD, BC, BD, and CD). Each rater was assigned one of the six pairs of events and asked to determine the relative evil of all possible pairs of situations within that pair of events. That included both intra-event pairs (two situations from a single event) and inter-event pairs (two situations from different events). There were a total of 148 situation pairs to compare.

Five distinct classes or groups of raters were used in the experiment. They were:

- *Human military experts*. Humans with specific training or experience, gained in active military service, that was deemed to qualify them as experts.
- *Human nonmilitary experts*. Humans with specific training or experience, such backgrounds in ethics, religion, psychology, sociology, and political science, which was deemed to qualify them as experts.
- *Human nonexperts.* Humans without any relevant training or experience.
- *Random raters.* Computer random number generators that assigned ratings at random, with no consideration whatsoever of the actual situations being compared.
- *Models.* The two versions of the automated models that were being validated.

Each human rater was assigned to a single rater class, even though some of the raters were qualified as both military and nonmilitary experts. The information used to classify the human raters was provided by the raters. The validation experiment had 20 human military experts, 15 human nonmilitary experts, 15 human nonexperts, 12 random, and four models. The difference between the expert raters and the nonexpert raters is important; the experts' assessment of relative evil were assumed to be correct, and the various raters were evaluated based on their agreement with the experts.

Each of the six event pairs (AB, AC, AD, BC, BD, CD) was assigned at least two and as many as five human raters in each of the three human categories. The differences in the number of raters assigned to an event pair was an artifact of the variability of response success and response time experienced from the raters. To avoid order bias in the surveys, each of the six event pair surveys was assembled in two versions, with the events described in reverse order between the two. Approximately half of the human raters were given each version. Table 3 presents the total number of human raters who responded with usable ratings for each event pair. Each human rater had from 10 to 15 situation pairs to rate.

For a given situation pair, the rater was asked to compare the evil associated with the first situation (e.g., situation A1, the Germans in the Warsaw uprising) to the evil associated with the second situation (e.g., situation B2, the Soviets in the Korean Air Lines Flight 007 shoot down). The rater rated the relative evil of the two situations using a Likert scale type response (Likert, 1932). This scale was expected to produce more reliable results than asking the raters to provide absolute numerical ratings. For this experiment, evil was defined roughly and instructions provided as follows:

> *You are requested to read the information on the first event, then compare the relative "evil" associated with certain parties' actions during that event from your own perspective. For the purposes of this experiment, "evil" is defined to be intentional or anticipatable harm resulting from a given action.*

In addition to the 50 human raters, there were 12 random raters and two model raters. Each of these raters rated all six of the event pairs. These raters by design produced absolute numerical ratings of the evil associated with a situation. Because they produced absolute numerical ratings, it was necessary to convert those ratings to the same Likert scale relative responses produced by the human raters so that the nonhuman and human ratings could be compared. This conversion was done using a procedure that compared the ratings, and based on their difference, produced a Likert scale response.

Measuring rater performance

Measuring the performance of each rater involved calculating an agreement rating for the rater, then rank ordering the set of all raters with respect to their agreement ratings. The ranking of each rater and the average ranking of each class of raters provides a measurement of how well each rater and each class, respectively, agrees with the experts. Not every expert will agree with the others on the assessment of every situation pair, and certain experts may disagree with the set of all experts much more than others. Therefore, the individual experts were included in the rank ordering according to their agreement ratings.

The agreement rating itself is a measure of how well a rater's assessments agree with the assessments of experts. Human raters assess only a subset of the situation pairs—those involving the two events in the packet that he or she receives. Therefore, only the situation pairs evaluated by the rater are taken into account when calculating their agreement rating. For each situation pair where the rater has provided a response, and for each expert who has also provided a response for that situation pair, a comparison is made. The agreement rating for the rater represents the number of comparisons for which the rater agrees with an expert. Experts' ratings are only compared to the other experts' ratings—that is, they are not compared to themselves.

CALIBRATION PROCESS, RESULTS, FINDINGS, AND LIMITATIONS

The calibration process involves adjusting the input parameter weights and other factors that the models use so that they are more closely aligned with expert assessments. This section reports the calibration process, results, and findings of the project. First, the process of calibrating each model is discussed. Then, the numerical results of the calibration, the results of the ratings, and the degree of agreement between the expert raters and the other classes of raters are reported. Following that, limitations on the results due to certain aspects of the experiment and the rater pool are listed. Finally, the results are interpreted in the context of the limitations to produce the findings of the model development and experiment.

Calibration process

The following describes the process of calibrating both Metric v0 and Metric v1 based upon expert data collected in the preceding experiment.

Metric v0 calibration process. To calibrate the initial model, a single parameter was devised—the increment between successive order of evil exponents, *I*. To determine the optimal value for *I* in Metric v0, the parameter was adjusted via hill-climbing optimization. The default value for *I* is 1. In terms of *I*, the values for the order of evil exponents become 1 for necessary, $1 + I$ for consequential, $1 + 2I$ for selfish, and $1 + 3I$ for malicious. As *I* is modified, the model's scale factors are affected, which in turn affects how the model rates each situation pair. Thus, the model's ranking may increase or decrease as *I* is adjusted. For example, a lower *I* decreases the overall effect that, for example, the ''maliciousness'' of a situation affects its overall evil rating; a higher *I* increases the effect. Several values for *I* were attempted methodically. Starting at $I = 0.125$ produced a rank of 47 for the model. This steadily climbed as *I* became larger, finally achieving a rank of 30 at $I = 6.0$. The rank stagnated with further increases of *I*.

Metric v1 calibration process. In total, Metric v1 contains more than 30 adjustable parameters—a computational search across the entire parameter space for a set of optimum values is untenable. Thus, the chosen calibration process for Metric v1 was an expert-guided heuristic search, an understanding of human psychology and moral factors, and an understanding of the historical events themselves. The parameter weights incrementally adjusted, then the model was rerun using those weights to produce a ranking for the model. At a particular point, no further adjustments could increase the ranking of the model.

Results

Regarding Metric v0, it is apparent that 30 is the optimum ranking for Metric v1—no further increase of *I* yields better results. Two different weighting schemes for Metric v1 were used: unskewed and skewed. The unskewed weights represent an attempt to keep all input parameters intact in the model. In contrast, the set of skewed weights makes no such attempt; in this weighting scheme, input parameters' weights could potentially become 0 percent, thus potentially removing certain input parameters from consideration. Both weighting schemes were calibrated based upon expert assessments. Thus, including Metric v0, there are four different model formulations represented in the rankings: the original Metric v0, the calibrated version of that metric, the unskewed version of Metric v1, and the skewed version of Metric v1. It was found that the unskewed weightings for Metric v1 presented significant issues with two particular situation pairs: the Soviets in the KAL 007 shoot down event and the Cuban communists in the Bay of Pigs event. Thus, rankings were also calculated without considering any assessments that involved these two situation pairs, although the weightings were calibrated with all situation pairs in mind. Moreover, in the skewed version of Metric v1, noncombatant casualties calibrate to a weight of 90 percent, and the diminishment factor is less than 1.0.

Table 4 shows the full results of the calibration process. The following conclusions can be drawn from this assessment:

- With its skewed weightings, Metric v1's ranking is comparable to that of all experts.
- Nonmilitary ethicists tend to agree with the set of all experts more than any other class, which separates them from both military experts and nonexperts. This may be a reflection of their extensive training in fields related to assessing evil and other moral issues.
- Military experts disagree more than all other classes of human raters, including nonexperts.

- Human raters and the models all provided meaningful assessments; at the very least, they all performed significantly better than the random evaluators.
- There is a significant improvement for calibrated versions of Metric v1 and Metric v0.
- Removing the ''problematic'' situation pairs from analysis increases the unskewed model's rankings, but decreases the skewed model's rankings. This indicates that the formulation that uses skewed weightings incorporates these two situation pairs more appropriately.
- In assessing the optimal calibrated weightings of both versions of the model, it appears that the presence of noncombatant deaths, especially if malicious, is overwhelmingly the deciding factor that experts tend to use—consciously or not—to assess situation pairs, although there may be other subtleties to their assessments that are not captured by the models.

The disagreement among military experts (and the agreement among nonmilitary experts) is worth noting. It hints that the model may be more readily calibrated to nonmilitary expert opinion. Due to the relatively small sample size associated with this experiment, however, the set of experts could not be split in order to compare or calibrate the model to the assessment of only the nonmilitary experts. Additional findings were as follows:

- Among human raters, the most disagreed-upon comparisons were, in order: USA/CIA actions in Bay of Pigs to the Taliban actions in Operation Enduring Freedom; Cuban exiles' actions in the Bay of Pigs to the Taliban actions in Operation Enduring Freedom; USA/CIA actions in Bay of Pigs to the German actions in the Warsaw Uprising. Thus, overall, the Bay of Pigs situations and the situation involving the Taliban were the most controversial.
- In contrast, the clearest comparisons were: Soviet actions in the KAL 007 shoot down seen as more evil than the Polish actions in the Warsaw Uprising; Taliban actions in Operation Enduring Freedom seen as more evil than the non-Soviet actions in the KAL 007 shoot down.

Limitations on results

Although this method presents potential for quantifying the nature of evil and providing military commanders with an ethical sense of their COAs, there are some limitations on the specific results. Most of these limitations involve the validation experiment—the process of gathering raters' assessments, the number of assessments gathered, the means through which assessments were gathered, the backgrounds and nature of the raters, and calibrating the model based upon rater assessments.

Indecision in comparisons. A few raters remarked that they could not decide on any situation pair comparison. This is primarily due to the definition of evil used. Some stated that they could not work with the definition of evil provided. It was apparent, however, that some raters legitimately believed that all comparisons were equal based upon the definition of evil provided, and recorded their responses accordingly. For the latter case, the models' weights and overall designs were refined with these responses intact, incorporating the legitimate beliefs of raters who could not decide between situation pairs.

Another reason some raters cited as why they could not make comparisons was that many inter-event situation pairs were vastly different. Although it is difficult to compare certain situations, this still warrants some attention; after all, in the course of their own analysis, military commanders are forced to decide between two COAs while keeping several variables in mind.

Despite the understandable lack of certain comparisons, however, the model would be more useful if more actual comparisons between events could be made. This may be partially corrected with an enhanced validation experiment.

Baseline morality. Another limitation on the result of the validation experiment is that only one baseline morality is captured—that which incorporates American values. Although this was the intent of the validation experiment because the target audience for the model is American military commanders, it must be noted that this limits the cultural perspectives through which the ethical implications of actions may be considered. Cultures aside from American cultures may incorporate completely different moral systems.

The framework theoretically may incorporate different baseline moralities; due to its flexibility in terms of input parameter weights and factors such as how much intention matters in a given culture, it is likely flexible enough to incorporate different baseline moralities.

Sample size. The research team contacted nearly 350 individuals, including approximately 300 experts, and received around 75 positive responses overall and around 50 positive responses from experts. Only a subset of these positive responses—albeit a large subset—actually submitted questionnaire packets. A larger sample size would enhance the results in terms of its statistical validity. More repetitions of each individual questionnaire packet would lend more credence to the validation experiment overall.

Findings

As shown, two versions of the model, Metric v1 with skewed weights and the calibrated Metric v0, agreed with the human expert raters in terms of average rank and average agreement over all situations approximately as well as each of the three groups of human raters (military experts, nonmilitary experts, and nonexperts) did individually. In fact, the average rank score for the model is better than all three groups of human raters, and the average agreement score for the model is better than one of the three groups of human raters, but these scores are all quite close and the differences between the three human groups' scores and the model are generally within the margin of error in the experiment. The best characterization is that the three groups of human raters and the model are closely comparable, which is a satisfying outcome given the newness of the model.

The results suggest that the model's formulas and weights have largely captured the essence, in a quantitative way, of the criteria and sensibilities used by the human experts in assessing the evil associated with military courses of action, and that the model's primary objective representing those experts with a computational tool is achievable.

- *Finding 1.* The concept of a quantitative assessment of the evil associated with a military COA is viable and practical. The models have many input variables in several different categories that include both human and infrastructure consequences of the COA. These inputs form an effective checklist of items to consider when comparing courses of action, and the appropriateness of the checklist is confirmed by the results. The effort required to prepare those inputs, while certainly not negligible, is likely to provide the analyst an understanding of the consequences of a COA that has value in and of itself beyond the quantification the model computes.
- *Finding 2.* Thoughtful use of a ''Metric of 'Evil''' of the type developed and tested can provide a valuable secondary benefit in terms of the analyst's understanding of the possible consequences of a COA. Although the model in its current form performed well in the validation experiment, it is nevertheless still laboratory software in need of further development, testing, and productizing.
- *Finding 3.* Before the model could be fielded and used operationally, a number of improvements and enhancements are required.

FUTURE WORK

Potential future work may address current limitations of the model and to allow it to more realistically or holistically represent ethical concerns. The model's design can handle an arbitrary number of input variables, and additions to and refinements of the input variable list have been identified. Proposed additions include harm to animals and damage to the environment. An important refinement is determination of the best unit of measurement for each of the different variables; it is not clear, for example, that the current model's simple counts of treaty violations and infrastructure damage events are the best choice.

Improvements to the model's representation of context and intent are possible. The definition, and thus the measurement and comparison, of ''evil'' seems to hinge on context and intent. Killing an enemy combatant is generally considered acceptable or not depending on whether or not he has surrendered. Better definitions of intentional harm and when intentional harm is evil are needed. Since completion of the work outlined here, we have further explored potential expansions of the model's philosophical context (see Reed, 2012).

The inability of some raters to compare the evil in certain situation pairs could be mitigated with a clarified definition of evil, or the term's replacement with something else, as well as an improved set of situations with more availability of comparable pairs (i.e., fewer situation pairs that were vastly different in scope or context). A larger experiment with an increased number and broader spectrum of raters, expert and nonexpert, would increase confidence and statistical power in the analysis. A repeat of the experiments with raters from a culture or baseline morality other than American would potentially be very revealing.

ACKNOWLEDGEMENTS

This research was supported by the US Army Aviation and Missile Research and Development and Engineering Center's System Simulation and Development Directorate under contract W31P4Q-09-D-0010. All views and opinions expressed are those of the authors and do not necessarily represent the views of the Department of the Army. Matthew J. Menarchek assisted with the historical research, and Joshua M. Ciardelli assisted with the model implementation. That support is gratefully acknowledged.

REFERENCES

Babbie, E. 2002. *The Basics of Social Research*, 2nd ed., Wadsworth Group.

Barbosa, S. E., and Petty, M. D. 2010. A Survey and Comparison of Past Instances of Combat Model Validation by Retrodiction, *Proceedings of the Spring 2010 Simulation Interoperability Workshop.*

Batschelet, A. 2002. Effects-Based Operations: A new operational model? US Army War College.

Baumeister, R. F. 1999. *Evil*, Henry Holt and Company.

Egan, C., and Reaper, J. 2007. Course of Action Scoring and Analysis, *Proceedings of the International Symposium on Collaborative Technologies and Systems* (ICCRTS), IEEE, pp. 297–303.

Fallesen, J. 1995. Decision Matrices and Time in Tactical Course of Action Analysis, *Military Psychology*, Vol 7 No 1, pp. 39–51.

FM 101-5. 1997. *Staff Organization and Operations.* Department of the Army.

Goodwin, P., and Wright, G. 2004. *Decision Analysis for Management Judgment*, 3rd ed., John Wiley & Sons.

Hanna, J., Reaper, J., Cox, T., and Walter, M. 2004. Course of Action Simulation Analysis, Command and Control Research and Technology Symposium: The Future of C2.

Heuer, R. 1999. *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, Central Intelligence Agency.

Illustrated Oxford Dictionary. 2003. DK Publishing, Inc.

Likert, R. 1932. A Technique for the Measurement of Attitudes, *Archives of Psychology*, Vol 22, No 140, pp. 1–55.

Miller, A. G. 2004. Introduction and Overview, *The Social Psychology of Good and Evil*, A. G. Miller, ed., The Guilford Press.

Reed, G. S., and Jones, N. J. 2012. Toward Modeling and Automating Ethical Decision-Making: Design, Implementation, Limitations, and Responsibilities, *Topoi: An International Review of Philosophy*, Vol 32, No 2, pp. 237–250; doi: 10.1007/s11245-012-9127-x.

Richardson, D., Deckro, F., and Wiley, V. 2004. Modeling and Analysis of Post-Conflict Reconstruction, *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* (JDMS), Vol 1, No 4, pp. 201–213.

Staub, E. 2004. Basic Human Needs, Altruism, and Aggression, *The Social Psychology of Good and Evil*, A. G. Miller, ed., The Guilford Press.

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

Welner, M. 2006. Classifying Crimes by Severity: From Aggravators to Depravity, *A Crime Classification Manual*, J. Douglass et al., eds., Jossey-Bass.

Zimbardo, P. G. 2004. A Situationist Perspective on the Psychology of Evil: Understanding How Good People are Transformed into Perpetrators, *The Social Psychology of Good and Evil*, A. G. Miller, ed., The Guilford Press.

Zimbardo, P. G. 2007. *The Lucifer Effect*, Random House.