Journal of Military Ethics

ISSN: 1502-7570 (Print) 1502-7589 (Online) Journal homepage: www.tandfonline.com/journals/smil20

Supporting Ethical Decision-Making for Lethal Autonomous Weapons

Spencer Kohn, Marvin Cohen, Athena Johnson, Mikhail Terman, Gershon Weltman & Joseph Lyons

To cite this article: Spencer Kohn, Marvin Cohen, Athena Johnson, Mikhail Terman, Gershon Weltman & Joseph Lyons (2024) Supporting Ethical Decision-Making for Lethal Autonomous Weapons, Journal of Military Ethics, 23:1, 12-31, DOI: 10.1080/15027570.2024.2366094

To link to this article: https://doi.org/10.1080/15027570.2024.2366094

Published online: 25 Jun 2024.

JOURNAL OF MILITARY ETHICS
2024, VOL. 23, NO. 1, 12–31
https://doi.org/10.1080/15027570.2024.2366094

Routledge
Taylor & Francis Group

Supporting Ethical Decision-Making for Lethal Autonomous Weapons

Spencer Kohn[a], Marvin Cohen[a], Athena Johnson[a], Mikhail Terman[a], Gershon Weltman[a], and Joseph Lyons[b]

[a]Research and Development Department, Perceptronics Solutions, Inc., Sherman Oaks, CA, USA; [b]Air Force Research Laboratory, Dayton, OH, USA

ABSTRACT

This article describes a new and innovative methodology for calibrating trust in ethical actions by Lethal Autonomous Weapon Systems (LAWS). For the foreseeable future, LAWS will require human operators for mission planning, decision-making, and supervisory control; yet humans lack the cognitive bandwidth and processing speed to make prompt, real-time ethical decisions. As a result, trustworthy Artificial Intelligence (AI) will be required to support ethical decision-making. We use a Bayesian ethical decision model for: (1) human setting of ethical preferences and thresholds in accordance with Laws of War and tactical criteria; and (2) highlighting the factors that contribute to strike/no-strike recommendations for human evaluation. The model can perform an ethical analysis, provide a quantitative ethical strike/no-strike score, and recommend actions to reduce decision uncertainty. In this article, we describe successful initial evaluation trials of the Bayesian model and of a human interface for interaction with the model. Our Bayesian ethical decision model has an immediate application in wargames; the model can also be used to train operators in understanding the principles and key factors relevant to ethical decision-making; and it may eventually be used in actual military operations employing LAWS.

KEYWORDS

Ethical AI; Lethal autonomous weapons; Bayesian modeling

Introduction: artificial intelligence (AI) and human decision-making

For lethal intelligent weapons systems ever to be acceptable they will require the input of commanders, analysts and operators for mission planning, operational decision-making, and supervisory control. However, humans do not always have the cognitive bandwidth or processing speed to maintain high situational awareness in dynamic ethical dilemmas when many mission tasks are demanded concurrently. Artificial Intelligence has the capacity to process multiple information streams simultaneously, but its real-time ethical capabilities are limited. Ideally, the human and AI would operate as a team

CONTACT Gershon Weltman gweltman@percsolutions.com Perceptronics Solutions, Inc., 3527 Beverly Glen Blvd., Sherman Oaks, CA, USA

This article has been republished with minor changes. These changes do not impact the academic content of the article.

utilizing their relative strengths. In that case, the critical question becomes: *Can the operator trust the AI to understand the factors that shape ethical decisions and highlight the factors that shape ethically appropriate actions in the context of the Laws of War as the commander intends?* Recent DoD directives such as "Autonomy in Weapon Systems" (DoD 2023) govern the deployment of lethal and non-lethal force by autonomous or semi-autonomous weapon systems. The directive dictates that the identified systems will allow "commanders and operators *to exercise appropriate levels of human judgment over the use of force."* Accordingly, future military training and planning, including wargames, will require a new consideration of lethal intelligent weapons systems as part of multi-domain operations.

This article describes a modeling approach to highlight how different factors within the Laws of War impact the ethicality of a strike/no-strike action by a lethal autonomous weapon. Our methodology uses a Bayesian ethical decision model for: (1) Human setting of ethical preferences and thresholds in accordance with Laws of War and tactical criteria; and (2) Highlighting the dynamic impacts of the model inputs on strike/no-strike actions. These goals are achieved by incorporating data about a threat into the model, performing an ethical analysis based on pre-defined ethical preferences, recommending actions that could reduce uncertainty, and providing an ethics-based strike/no-strike score based on the current inputs. The methodology can be used to teach commanders, analysts, and operators the principles of ethical decision-making, and to enable comparative analysis of their choices. In the long-term an empirically validated version of this model may be used to speed and/or augment ethical decision-making during actual mission execution.

Autonomy and ethical trust

Decades of research have focused on how much to automate and when (Parasuraman, Sheridan, and Wickens 2000). Recent research has highlighted the importance of mixed human-autonomy relationships that focus on unique collaborations in the context of personal robotics and unmanned systems (Onnasch et al. 2014). The ability to appropriately trust automated aids is essential for operators (Lee and See 2004), especially under the high-risk, extremely vulnerable conditions in which lethal intelligent weapons systems will be used (Shanahan 2016). An operator who trusts the aid in the situations where its recommendations are appropriate will have much more efficient and accurate outcomes compared to manually performing the task themselves. Trust in the current case is defined as performance trust (Malle and Ullman 2021), which in this context is the capability of the system to provide appropriate decisions within an ethical framework provided by humans. In the performance-trust conceptualization, an operator's performance trust is well calibrated if their subjective trust and behavioral compliance match the operational trustworthiness of the automation.

Perceptronics Solutions, with which the authors of this article are affiliated, has previously described a solution for calibration of operational trust in autonomous systems that includes trust cues for transparent and explainable autonomy that moves the operator to a calibrated trust condition (de Visser et al. 2014; Freedy et al. 2007). The new "ethical trust" situation for lethal autonomous weapons is analogous to performance trust for the general case of autonomous systems. If the operator of a lethal autonomous weapon system over-trusts, the system's actions may violate ethical principles with severe consequences; if

the operator under-trusts, the lethal autonomous weapon system may fail to act at all when needed for combat (Bahner et al. 2008). The prior methodology has been newly adapted to allow proper calibration of the ethical capabilities of lethal autonomous weapons.

Bayesian decision modeling

Bayesian decision models possess features that are critical for successful guidance in ambiguous and dynamic ethical scenarios. In our implementation, a network of nodes represents the connection of causes to effects along with the influence of those effects on ethical judgment. Uncertain causes and effects are assigned probabilities, which are based initially on prior information about conditions and events in the situation and then updated as new information comes in from any source. The information is propagated through the network to update expectations for action outcomes, and possible outcomes are then filtered through importance weights based on Laws of War, rules of engagement, commander's intent, and tactical exigencies. The final model output is a number representing the degree of ethical permissibility or impermissibility of the action under consideration and an explanation (if desired) of factors that were critical in that determination.

This approach differs in many respects from more rigid traditional modeling, which tends to process a less comprehensive set of variables by means of fixed parameters and less coherent algorithms. The result is that Bayesian models are more flexible and robust under conditions that combine uncertainty, rapid change, and ethical nuance. Devitt (2021) laid out a set of constraints for an effective "Bayesian epistemology" in the context of LAWS. Our Bayesian ethical decision model satisfies those constraints. It can:

(1) Distinguish situational factors from outcomes and make their causal relationships explicit in visual graphs;
(2) Represent graded beliefs by the probabilities of uncertain states of situation factors and outcomes;
(3) Enable initial beliefs about situation factors and outcomes to evolve over time, e.g. become more or less certain as new observations or intelligence reports clarify the situation;
(4) Represent ethical implications and degrees of preference for outcome states by assigned weights and calculated values;
(5) Allow final ethical permissibility scores to be intelligibly interpreted as the expected utility of alternative options for action.

Bayesian modeling thus provides a new and uniquely powerful way of representing the ethical constituents of a LAWS Strike/No-Strike decision: by means of a computable model that incorporates probabilistic representations of situation variables, the expected outcomes of striking or not striking, and outcome values that allow a comparison of options that is both rational and context sensitive. The result is more nuanced and transparently defensible support for an "Ethically Correct Decision."

Legal expert Alan Dershowitz offers an additional ethical argument in favor of explicitly recognizing *all* considerations that should enter into LAWS decision-making, rather than excessively simplifying an ethical model for easy human comprehension (Dershowitz 2022). He posits that a democracy should not ask its citizens to play the role of

hard-bitten fictional detective Dirty Harry, i.e. take actions that violate its laws or rules, or commit actions that cannot be disclosed even though the actions were justified and necessary. The conditions under which such actions are justified (or prohibited) should be explicitly laid out in the laws and rules themselves, so that decisions can be openly explained and defended. Dershowitz extends this argument to LAWS decision-making: "If robotic warfare is to be used – as it will be increasingly – it is imperative that the programs be as explicit as technically feasible. Indeed it may be a virtue of robotic "soldiers," who cannot exercise human discretion, that inevitable discretionary decisions will have to be made in anticipation of battlefield choices. And they will have to be made by humans based on hypothetical probabilities. In a democracy that is the least worst process" (Dershowitz 2022).

Material and methods

Calibration system concept

Figure 1 shows our innovative system concept for ethical trust calibration.

The main system components as numbered in Figure 1 are described in the following.

(1) LAWS Scenarios. These are a set of fictional scenarios involving potential autonomous UAV strikes. They are representative of current military missions, but do not include any actual mission planning data.
(2) Ethical Decision Model. This is a Bayesian model that incorporates both tactical and ethical factors for each strike scenario to provide an ethical score ranging from highly non-permissible to highly permissible.
(3) Ethical Rules. These are the ethical guidelines incorporated in the Bayesian model deriving from the widely accepted Laws of War and the Rules of Engagement for specific scenarios.
(4) Operator Calibration. Ethical trust calibration of the LAWS operators reflects whether the operators (and commanders and analysts) accurately trust the

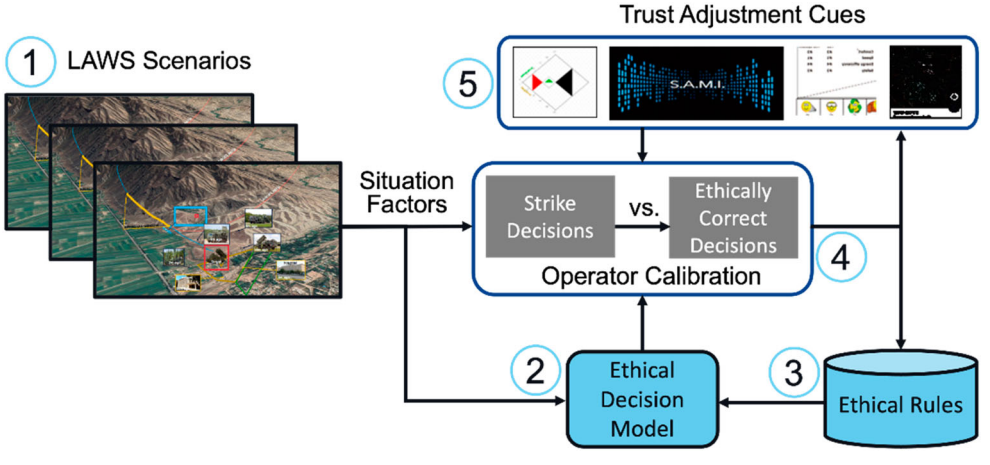

Figure 1. System concept for calibration of trust in ethical decision-making.

model's ability to capture and appropriately weight the factors in the strike situations.

(5) Calibration Cues. Where operator trust adjustment is required, previously validated trust adjustment cues can be used to get the operator properly calibrated. And in some cases, operator (and commander and analyst) responses based on experience and analysis will identify necessary modifications in the Ethical Rules.

LAWS scenarios

We have adopted for our general scenario the case where a target has been surveilled by a drone to obtain coverage information which is reconciled with prior information about the area as well as recent ATR (average true range) data. As shown in Figure 2, the autonomous UAVs have two main types of information: (1) Prior Information – What the system generally knows about the area and what is there; and (2) Dynamic Information – How long it has been since the system last received data about the area. What does the automatic target recognition indicate? The UAV weapon potentially affects three danger zones: (1) Extreme (orange) – Highly likely to involve damage from the strike; (2) High (yellow) – Very likely to involve damage from the strike; (3) Moderate (brown) – Less likely and/or unlikely to involve damage from the strike. Various installations are co-located in these zones, and their value and survivability are governed by the ethical framework underpinning the decision-making system.

Our immediate instantiation of this scenario is counter-insurgency (COIN), the ever-present battlefield of the last two decades, where hostile forces are emplaced near civilians, or practically indistinguishable from them. The importance of ethical decisions is clear when both civilian and allied forces are in danger. However, we are oriented towards the next battlefield per the recommendation of our SMEs (Subject Matter

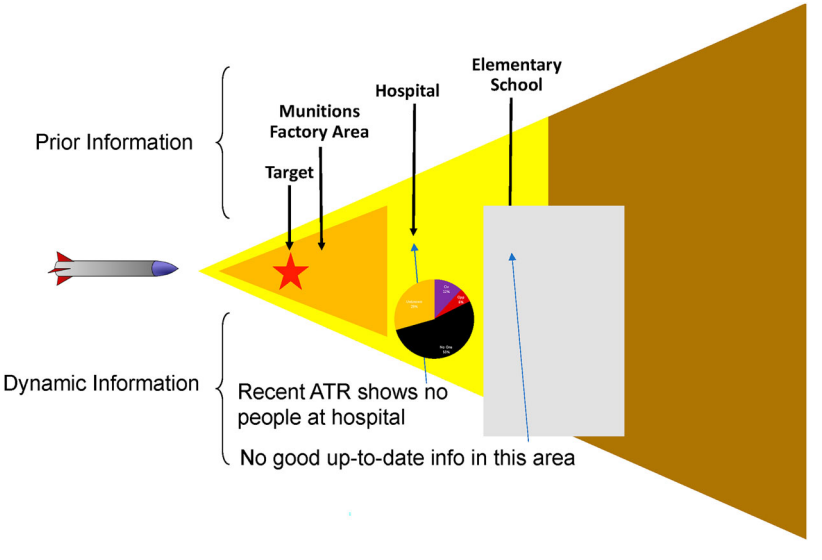

Figure 2. Basic LAWS scenario paradigm.

Experts) and interviewees – Eastern Europe or the South China Sea. The more ominous threat is in the South China Sea, where multiple nation states are laying claim to territorial waters. Ambiguous alliances between China, the Philippines, Indonesia, Malaysia, Vietnam, and allied nations, disguised auxiliary fishing fleets with EW sensors and possible kinetic weapons, and the possibility for hostile autonomous drone swarms could lead to an ethically murky engagement. Accordingly, this scenario may be an ideal test case for ethical LAWS, and a model or interface designed for them. Operators would need to make decisions in quick succession, requiring collaboration with systems that augment and accelerate their ethical decision-making.

Conformity with laws of war

An essential part of ethical decision modeling for LAWS is compliance with the Laws of War (Arkin 2010; Filkins 2021; Galliott 2021). The Laws are generally stated as requiring: (1) Military Necessity for the action of war; (2) Discrimination among combatants and non-combatants; (3) Avoidance of unnecessary suffering; (4) Proportionality in response to provocation; and (5) Honor in terms of fairness and mutual respect (DoD 2011).

However, the Laws of War can be ambiguous in concrete cases. The implications for action within a high-level framework depend on two somewhat subjective elements: how the situation is interpreted and the relative weight placed on different principles or frameworks. The deontological ethical framework proposes duties that are meant to be inviolable, e.g. do not intentionally attack innocent civilians (Arkin 2009). This is parallel to the principle of *Discrimination* for the conduct of war, which prohibits inflicting harm on non-combatants for no valid military purpose. However, neither of these formulations takes account of uncertainty. In other words, is attack permitted if there is *any* chance that the target involves enemy combatants? That policy accepts almost any level of risk, hence, it renders the principle of Discrimination powerless to spare civilians. Conversely, is attack forbidden if there is any chance that the target involves non-combatants? That policy might block reasonable applications of Military Necessity in combat decision-making.

One practical conclusion is that most decisions will be based on degrees of uncertainty between the two extremes of very small and very large risk. A second conclusion is that potentially competing principles (e.g. Military Necessity and Discrimination) must be weighted by their relative importance. The Bayesian model is ideally suited for both functions. As explained in the following section, the Bayesian model accommodates both the spirit and the letter of the law. But the model alone does not dictate answers to ethical problems; for specific cases it requires expert judgment in the interpretation of situations, and it allows flexibility in the prioritization of conflicting evaluative principles. Nevertheless, the Bayesian model can assure that the Laws of War and associated ethical principles are carefully and rationally accounted for in tactical decisions.

Measuring trust in ethical decision-making

We are interested in capturing operators' trust in the ability of the model to correctly process inputs, provide an accurate shoot/no-shoot assessment of the situation, and present the aggregated information to operators to augment their decision-making.

Our primary measures of trust are operator compliance and self-reported trust regarding model decisions in ethically relevant training scenarios. In our calibration studies, these measures are collected after the operator has reviewed the scenario but before model's recommendation has been provided, and after viewing the model's output and a detailed explanation (e.g. in after-action review).

Bayesian ethical decision model

Figure 3 depicts our Bayesian model for the ethical permissibility of engaging a target. The model is based on military targeting guidance (US Army 1993) as well as Laws of War considerations. Each node represents a variable, and links between nodes represent the influence of "parent" variables on the variables depicted as their "children." The model is organized both by levels of reasoning and by topics. Concentric rings correspond to phases of reasoning. They begin on the periphery with model inputs provided by a human, potentially working in conjunction with automated aids, who analyzes data and doctrine to determine what value is appropriate. These inputs are integrated into assessments, which then inform predicted engagement outcomes. The final product is the evaluation of predicted outcomes both overall and on ethically relevant dimensions (represented by nodes in the innermost circle). The model is also organized by topics, corresponding to "pie slices." Variables that support engagement by appeal to Military Necessity are represented by green nodes in the upper half of the model. Variables for the Military Necessity of Immediate Defense are in the upper left quadrant. Variables for the Military Necessity of Long-term Benefit are in the upper right quadrant. Model inputs will come from operation, intelligence, or geo-political analysts with expertise in the relevant environments.

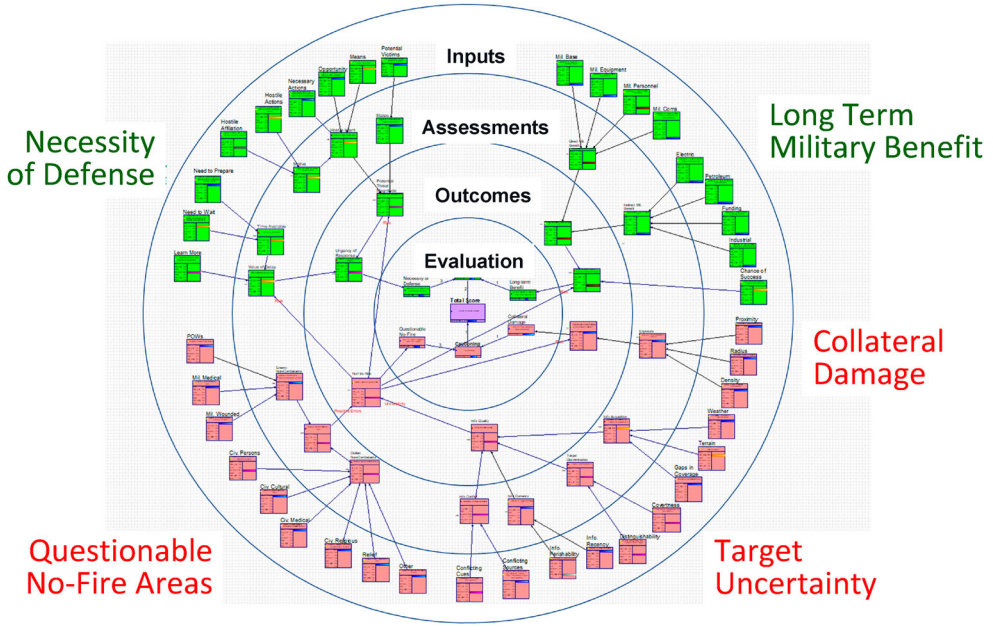

Figure 3. High-level visualization of Bayesian model for ethical LAWS decisions.

Figure 4 focuses on the "Necessity of Immediate Defense" sector of the model, with nodes representing current conditions in the scenario and possible action outcomes given those conditions. They are called *chance* nodes because probabilities are either assigned or calculated for their uncertain states. The focus in this part of the model is the effect of scenario elements on the magnitude and timing of a target threat, and the effect of the latter on ethical evaluation of an attack response. The key factors, as numbered in Figure 4, are related by probabilistic versions of AND and OR as follows: The Urgency of an attack response against the target depends on high potential threat magnitude AND low time availability AND low risk of a no-fire error. Potential threat magnitude depends on high probability of hostile intent AND a high level of damage it can cause. Time availability is low if the target is in position to attack *its* target AND we are in position to attack our target, OR nothing will be gained (no extra information or options) by a delay in our response.

Figure 5 depicts the central core of the model, which integrates inferences from all parts of the model into an ethical evaluation result. It also illustrates two additional node types in the Bayesian model and the connections between them. For example, Immediate Defense is a *utility* node, which assigns qualitative levels to outcomes (i.e. minimum, very low, low, medium, high, very high, and maximum) with corresponding numerical magnitudes. Probabilities of outcome levels (based on Figure 4) are combined with their magnitudes to generate a single numerical utility score for the Immediate

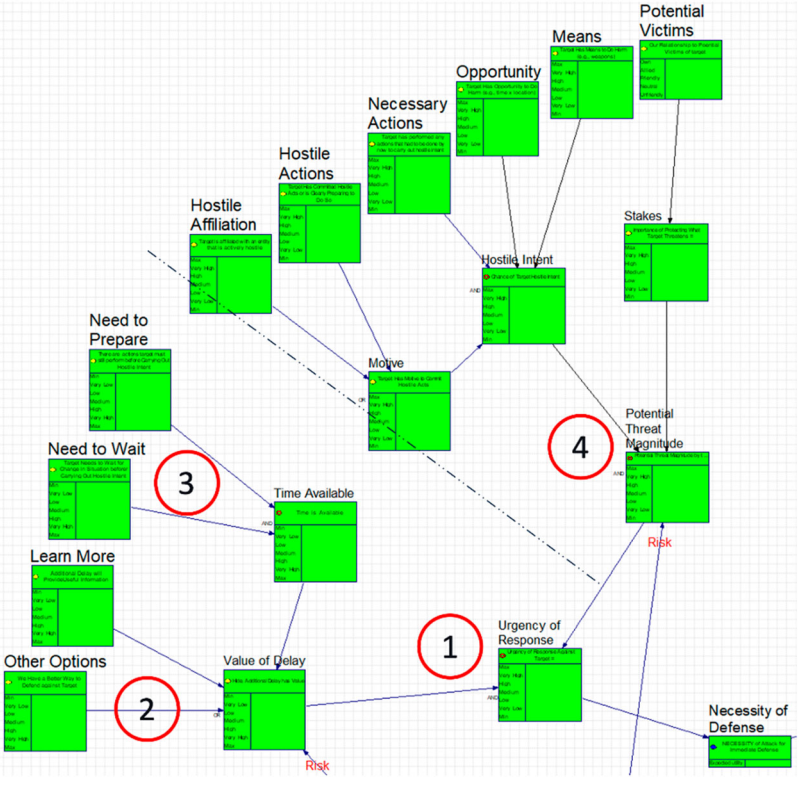

Figure 4. Necessity of defense section of the ethical decision model.

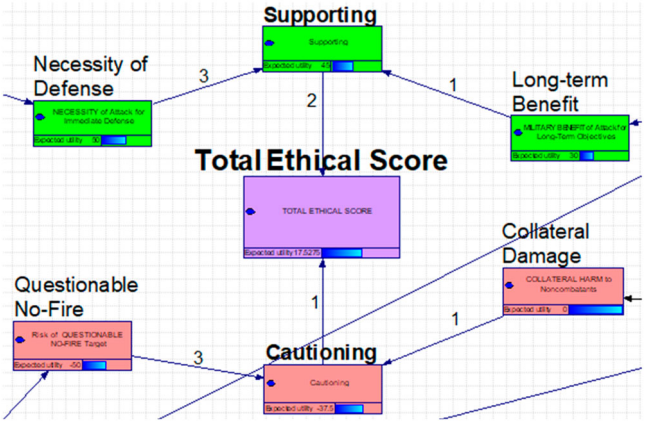

Figure 5. Model evaluation result.

Defense ethical evaluation node. Similar processes integrate inferences from other parts of the model, via utility nodes for Long-Term Benefit, No-Fire Discrimination, and Collateral Damage Proportionality.

Figure 5 further shows how the four ethical evaluation components are integrated. *Multi-attribute utility* nodes assign ethical importance weights to input utilities in order to combine them. Thus, the two parts of Military Necessity, i.e. Immediate Defense and Long-Term Benefit, are combined by a multi-attribute utility node representing positive Support for a strike. In parallel, the two negative evaluation elements, i.e. Questionable No-Fire Discrimination and Collateral Damage Proportionality, are combined by a multi-attribute utility node representing Ethical Caution. Each arrow in Figure 5 connecting utility nodes to a multi-attribute utility node is labeled with a number indicating the relative importance weights for the utility inputs. Finally, the two multi-attribute utility nodes – Support for a Strike and Caution Against a Strike – are combined via a higher-level multi-attribute utility node representing Total Ethical Permissibility.

In this example, Immediate Defense has three times the importance of Long-Term Benefit: protection of self or friendly assets receives more importance than contribution to a long-term objective (although they may both be in play). Similarly, Discrimination of Questionable No-Fire targets is three times more important than Collateral Damage: targeting a legitimate enemy asset is more important to the ethical evaluation than the potential of unintended collateral damage to nearby non-combatatants. Finally, factors in support of a strike have two times the importance of factors cautioning against it, reflecting current rules of engagement.

The formulae for integrating the top-level ethical components involve simple linear addition (after weights are normalized so that they add to 1.0):

$$\begin{aligned}&\text{Total Ethical Permissibility} \;=\\&(\text{Weight on Support for Attack})*(\text{Support for Attack})-\\&(\text{Weight on Ethical Caution})*(\text{Support for Ethical Caution})\end{aligned}$$

where

$$\begin{aligned}&\text{Support for Attack} \;=\\&(\text{Weight for Immediate Defense})*(\text{Necessity of Attack for Immediate Defense})\\&\quad+(\text{Weight for Long Term Military Benefit})*\text{Long-term Military Benefit of Attack})\end{aligned}$$

and

$$\begin{aligned}&\text{Support for Ethical Caution} \;=\\&(\text{Weight for Questionable No-Fire})*(\text{Risk of Questionable No-Fire})\\&\quad+(\text{Weight for Collateral Harm})*(\text{Risk of Collateral Harm})\end{aligned}$$

In the case of our basic LAWS scenario as described below, the result is a positive permissibility score of +17, which can be interpreted as a relatively weak recommendation to strike. Working with this network of competing weights, our sensitivity analysis allows us to identify which specific input nodes have the greatest influence on the ethical permissibility score.

Basic LAWS scenario

We adopted a basic LAWS scenario with variants for the purpose of exercising the ethical decision model and evaluating it with test operators. Our scenario is a fictionalized extrapolation of the accidental killing of journalists by an Apache helicopter pilot in Afghanistan in 2007 (Hodge 2010), with added elements of other similar incidents (Aikins 2021; Ponniah and Marinkovic 2019). The Afghanistan incident was exposed by Wikileaks and received negative international press. However, a military legal review verified that the attack was justified based on available information and rules of engagement. Our fictionalized permutation of that scenario involves a group of unknown armed people with unknown intent who are assembling at an intersection along the route of a friendly force convoy during a period of insurgency. It does not include any actual military mission planning data. It is unclear if the people are enemies, but they have proximity and means to attack friendly forces that are arriving by convoy in a few minutes. Our model considers that the adversaries have a high capability for covertness, but the density of target-similar distracters is also high – it is difficult to discern enemies from civilians. However, waiting for conclusive additional information via imagery would mean that the convoy has reached the nearby intersection and any attack would have already occurred. This scenario was used in both the model described above, as well as the calibration studies described throughout this article.

Interfacing with operators and commanders: training and wargaming

Our new Bayesian ethical decision model will facilitate ethically calibrated decision-making by targeting two different but complementary skills. First, the model will train operators to understand which ethically relevant decision points are most likely to affect outcomes. At multiple decision points during a mission, operators or commanders must make ethically relevant targeting decisions under time pressure. To perform

effectively, the operator must have appropriate situational awareness, and understand which decisions have the strongest effects on the ethicality of the outcome. The model's sensitivity analysis provides an ordered list of inputs that may affect the outcome, along with a brief explanation of considerations. Each input can be associated with available data (imagery, human intelligence reports, map data, etc.). By analyzing corresponding new data, the operator may either confirm the input state to increase the model's confidence in its score, or change the input value to affect the permissibility score. Repeated training in a variety of situations would help operators to internalize an understanding of which factors have an outsized impact on different scenarios, and would enable them to improve their own efficiency separate from this aid.

Second, comparative analysis tools built into the model help operators analyze the upstream effects of their choices. The user interface, shown later in Figure 8, compares the permissibility score as well as the scores of every affected model node for each operator choice. Operators can compare one decision path against another, evaluating the permissibility of an outcome if they changed their mission plan, or took the time to collect more data to uncover a hypothetical ground truth. Our aid again benefits training by building critical thinking skills during less time-sensitive conditions such as planning and after-action review. At a high-level these tools can be extended to wargaming, enabling commanders to understand whether their operators are making efficient and accurate decisions. Comparing the effects of team strategies on decision efficiency is crucial for the battlefields of the future, where swarms of autonomous UAVs and ALEs may be attacking and defending large swaths, such as in the South China Sea.

Results

Calibration test

We conducted an initial calibration test using interactive online materials and five test operators to help us determine whether representative operators' understanding of the LAWS' ability to take an ethically correct strike or no-strike action is consistent with the way the LAWS would behave in the immediate situation. We created five descriptive variations of the basic Apache strike scenario described above, with ethical scores ranging from – 33 (strongly cautioning against a strike) to +67 (strongly supporting a strike). Table 1 briefly describes the scenarios and shows the corresponding Ethical Score. The bias towards high positive scores reflects counter-insurgency operations and corresponding rules of engagement, yet it is important to reiterate that the scores are not an authorization to strike, no matter how high. Rather, they reflect the weight of evidence in the model, which the operator or commanders can use to assist their strike decision-making process.

Figure 6 (Left) compares the mean ethical scores assigned by the test operators for each scenario to the ethical scores calculated by the Bayesian decision model; the diagonal line represents a perfect equivalence between the two sets of scores. The results indicate a relatively close coherence from the neutral (−4) to the positive, or strike permissive, region (max +56), but a divergence in the negative, or strike caution, region (from – 14 to – 25), where the test operators assign more permissive scores than the model. Figure 6 (Right) shows the test operators' mean reported confidence

Table 1. Test scenario variants.

	Test scenario	Ethical score
1	Need to Wait; suspicious group has journalist markings and has not been aggressive	−25.0
2	Could be Journalists; intelligence says that journalists are operating in the area, and equipment could be cameras and microphones	−14.0
3	Credible Threat Decreasing; armed group remains under suspicion but is moving away from convoy	−4.0
4	Suspicious Gathering of Armed Men; armed group is assembling near route of convoy	+16.5
5	Confirmed Combatants; group of armed men on convoy route are confirmed combatants	+56.6

level for their own estimated score and for the model score assignment on a scale from 0 (no confidence at all) to 10 (high confidence) for each of the scenarios. Operators had consistently high confidence in the model's provided score, but inconsistent confidence in their own scores.

The alignment of operator and system scores for highly permissible scenarios is a positive sign for the calibration of our model, while the mis-calibration for low-permissibility scenarios is not unexpected. Clear threats are often self-evident to both the operator and model, while humans in a warzone may be prone to interpreting generally safe scenarios as being threatening. Consequently, there is a great need for the model to be extremely transparent and explain justification for its assessments in all scenarios. Transparency will help operators calibrate their assessments when utilizing the model, and will help calibrate their reliance on the model's assessment. Even without these transparency actions yet enacted, operators' confidence in the model is generally higher and more stable than their own self-confidence. Test operators recounted that the model inspired confidence because it was seen as collecting and evaluating more information with greater consistently than any single human. Combining the model's reach with transparency strategies explaining critical contributions to its ethical score may result in both better calibration and higher trust.

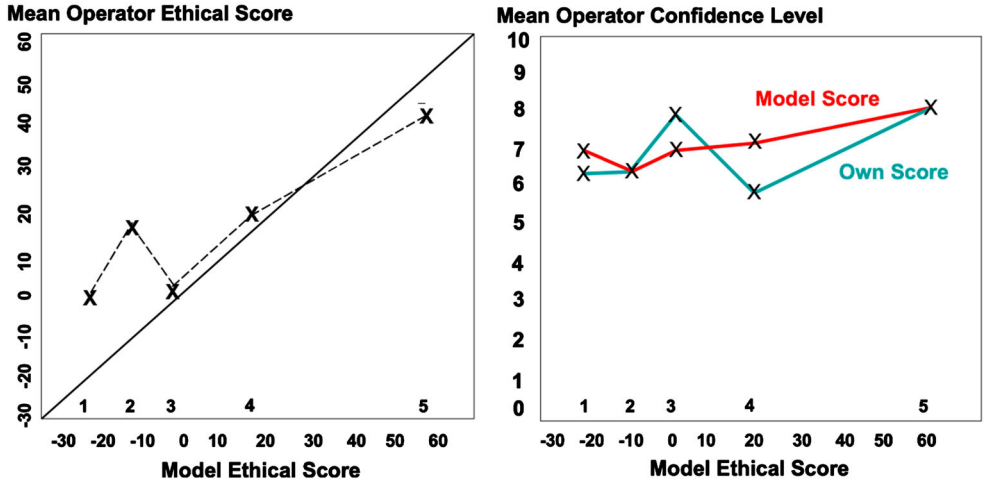

Figure 6. Operator ethical score (left) and confidence level (right) versus model ethical score.

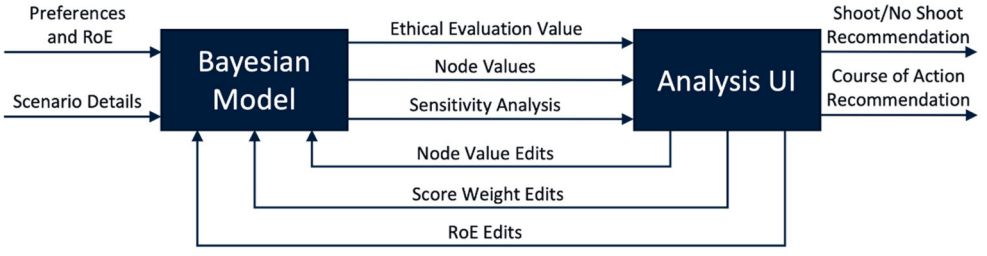

Figure 7. Data flow diagram.

User interface

We based our interface requirements on the high-level Data Flow Diagram (DFD) shown in Figure 7. We outlined the steps of the data exchange in accordance with the standard steps of the "kill chain" decision process employed in targeting activities. Our main workflow was: (1) View score (Shoot/no-shoot/uncertain); (2) Relay finding to Intel Analyst; (3) Wait until final plan is approved by Intel Analyst; and (4) Make changes according to Intel Analyst request to make the strike permissible *or* don't act if the strike is impermissible and cannot become permissible.

Figure 8 shows the prototype user interface configuration:

> The left side of the interface shows the nodes in the ethical decision model and the state of each node. The overall permissibility score is shown both at the heart of the model and above the model (−3.75 in this case, or barely not permissible). The right side of the interface initially features a list of the three most sensitive nodes in terms of their influence on the ethical decision score. Selecting "View" next to the node description provides the current state of the related node.

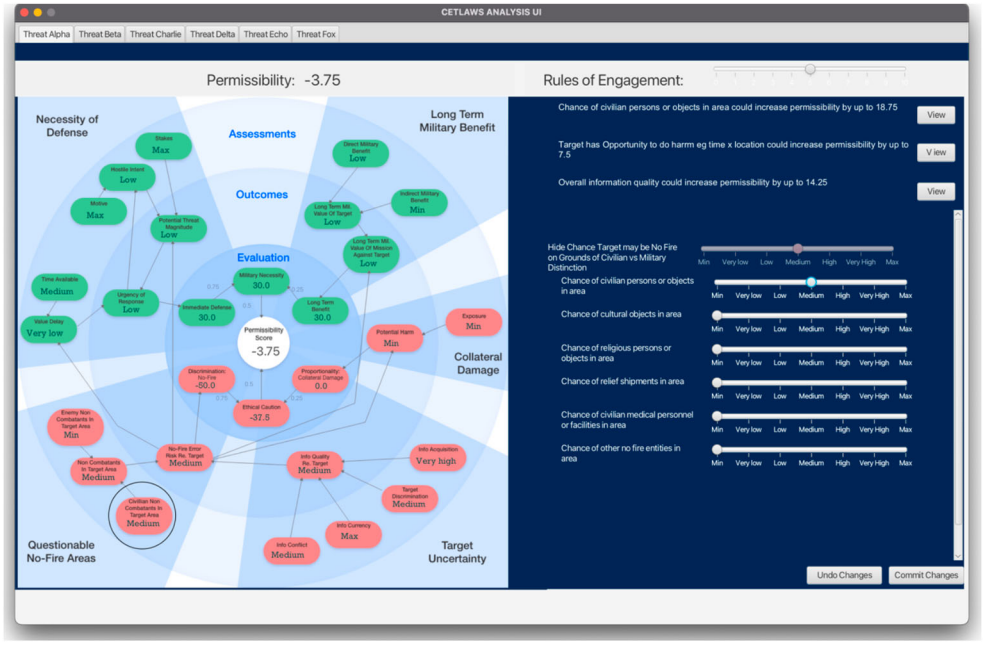

Figure 8. Prototype user interface.

In this example, the current state of the "Chance of civilian persons or objects in area" node is "Medium." Adjusting the sliders adjusts the state of the related input node and potentially higher-level nodes which are impacted by the change. For example, adjusting this slider down to "Very low," will produce several changes, that result in a change in overall score to +10.5. Choosing to "Undo Changes" resets the model to the initial state. Choosing "Commit Changes" instantiates the updated model with the new permissibility score of +10.5. This score generally indicates mild permissibility, but must be considered in context to determine whether that permissibility is high enough to continue the strike solution.

Interface evaluation

We performed an evaluation of the prototype interface using four former military personnel as representative commanders, analysts, or operators. Our main objectives were to: (1) Determine what cognitive workflow users would employ when evaluating ethical decisions; (2) Determine the degree to which participants find the model interface features useful; and (3) Evaluate the usability of the prototype features and identify future usability pain points or needed features. We elicited feedback on the workflow and features during interactive Zoom sessions focused on a sample scenario and the associated decision model outputs. We employed a semi-structured interview process, where questions focused on best practices within the military decision-making process, and on our prototype features and methods. The key evaluation responses are summarized in the following.

Our SMEs (Subject Matter Experts) appreciated orientation via the permissibility score and a concept that would display the active Rules of Engagement. They requested two specific improvements to speed ethical situational awareness: First, they requested thresholds for permissibility, such as a traffic light color-coding scheme, with hard boundaries on ethical scores that would enable them to quickly orient and ethically fire on the target. Second, several SMEs requested context on the confidence of the score, but they did not have a unified vision for implementation. Regardless of implementation, however, the SMEs said that reporting confidence in the current score would provide context on what actions needed to be performed.

Most of the SMEs had a positive evaluation of the model being shown on the left side of the user interface, declaring that they thought it would be useful after receiving some introductory training. The SMEs were all impressed with the input node update mechanism that allowed them to see the effects of potential changes to inputs. Once introduced to the sensitivity analysis tool, the SMEs were very enthusiastic about a tool that recommended a course of action to improve confidence in the model's estimate and to potentially increase permissibility. Similarly, all SMEs appreciated the sensitivity analysis' function allowing for the quick editing of relevant nodes. Most of the SMEs declared that the tool would be extremely helpful for analysts.

Feedback on team workflow generally confirmed the expected workflow while introducing some new potential interactions. According to the SMEs, the high-level user of this tool is the commander, who ultimately makes any strike decisions. The rank and position of the commander is dependent on the mission. In a dynamic self-defense scenario, the commander may have a lower rank and be physically close to the strike. However, the

SMEs reported that military doctrine is beginning to shift strike decision-making up the command chain. As result, commanders are more senior and responsible for a higher number of concurrent strike missions. Accordingly, a commander is unlikely to directly reference the model. Instead, they would rely on the permissibility score and summary reports delivered verbally. Current practice requires an analyst to answer commanders' questions and have a report ready to address potential courses of action. Given the relatively predictable nature of these questions, our SMEs recommended that the tool provide and update a written report to aid the analyst.

The analyst's role included updating inputs to reflect new data or alterations in the strike mission, enacting the actions proposed by the sensitivity analysis. Our SMEs reported that multiple specialized analysts may be simultaneously contributing to the model, with intelligence and operations analysts being responsible for different segments of the model. In this instance, it may be valuable to consider a version of the interface that is specialized to each analyst's role, where roles may be distinct in different regions and command structures. Regardless, at least one SME saw the value in a holistic model made available for decision-makers. Finally, our SMEs confirmed the need for commanders to trust the output of the system. According to two SMEs, trust is a product of absolute reliability in every situation, i.e. "performance trust". Upon prompting, the SMEs confirmed that transparency into the model's operation and uncertainties would increase trust in its ethical evaluation. Overall, the interviews emphasized the need for a trust focus throughout the development process. The system must be transparent and demonstrate a clear benefit to the decision-making process at every level, from the analysts to the commanders and leadership who may be acquiring this system.

Real-world test case

We conducted another evaluation of the model by examining if its ethical score would match the accepted action in a fictional version of a real-world strike/no-strike case: would the model's evaluations of ethical permissibility correspond to the observed judgments and decisions of highly experienced military professionals? We took our real-world case from Shortland, Alison, and Moran's book *Conflict: How Soldiers Make Impossible Decisions* (2019). Their examples are based on in-depth interviews with drone and manned aircraft pilots who served in Iraq or Afghanistan. The following is a typical case, but does not include any actual military planning information.

A drone pilot was given a counter-IED mission to investigate reports of people "burying stuff over by a main highway." If they were burying IEDs, the airman was allowed to use deadly force, i.e. to launch a missile strike at their location. The evidence could fit a persuasive story to that effect, yet the drone pilot decided not to engage. The relevant facts about this decision, as reported by Shortland, Alison, and Moran are: (1) "It did not take … long to see people digging and burying … . knew people were burying something because of temperature differences between the ground and the regular soil … 'people who, potentially, were planting IEDs in the ground;'" (2) " … could not get a clear picture as to what it was … from thousands of feet up in the air and through the haze and dirt … Even when they saw a motorcycle and a car drop something off,

'you couldn't see what it was. It didn't say IED on it. It was a very blurry picture;'" (3) Local civilians in his mind 'did some pretty whacky stuff on the highway … People do things completely different to [our expectations of social behaviors] … 'innocent people doing something random.'"

These few inputs, combined with several other plausible assumptions, were sufficient for our ethical decision model to recommend against attacking: The model assigned an ethical score of – 14. The model's recommendation matched the decision made by the real drone pilot, and it was made for similar reasons of information uncertainty and risk of error. Despite apparently persuasive evidence pointing in the opposite direction, the decision not to engage turned out to be correct. That evidence might have led a less astute pilot to mistakenly engage innocent farmers digging.

Discussion

Based on the calibration tests and the interface evaluation, we believe that our new Bayesian ethical decision model's features, workflow, and interface concept will be very well received by military personnel working with LAWS. One SME remarked "[the system] needs to be in every officer candidate school and NCO school." The feedback we received in the evaluation sessions makes our vision of calibrated Bayesian ethical decision modeling for LAWS wargaming and training operators practically achievable. Equally important, we received direction on how to better integrate the interface features into the existing military workflow to facilitate effective high-speed ethical decision-making by commanders. We believe that our real-world example, in which the model's recommendation matched the judgment of an experienced pilot, is representative of results to be obtained when our new Bayesian ethical model's recommendations are compared to real-world decisions for which there is consensus on the right choice – even though the model was not developed with those incidents in mind.

For its immediate application to wargames, our Bayesian ethical decision model will provide the following capabilities: (1) Helping wargame developers answer questions about the relevance and import of ethical decision factors in future wargames; (2) Determining the probability of success in terms of mission completion and satisfaction of ethical rules in ongoing wargames; (3) Calibrating wargame participants' trust in the ethical capabilities of decision aids and specific intelligent lethal weapon systems; and (4) Training the operators to understand the ethical decision-making principles and factors leading to key operational decisions.

In future operational applications the novel Bayesian ethical decision modeling features can be introduced in two important instances: (1) Setting the on-board ethical rules by which the LAWS makes its strike/no-strike decision to support enhanced understanding and trust by human operators and (potentially) for delegating bounded authority to LAWS prior to mission execution; and (2) Off-board ethical analysis used to determine whether the LAWS' onboard strike/no-strike decisions are ethically correct or incorrect as an adjunct to operator training, wargaming, and mission rehearsal. Our immediate use of Bayesian models will involve the following general cases for demonstration and evaluation: (1) Human setting of ethical preferences and thresholds/limits for the on-board LAWS models and off-board calibration models in accordance with

Laws of War, Rules of Engagement, situational variables, and other criteria; (2) Human in-the-loop interaction with the LAWS to interrogate and/or adjust the models as part of an operational mission; and perhaps in the future (3) Fully autonomous LAWS operation in which shoot/no-shoot decisions are made on the basis of pre-set values and without operator intervention. In all cases, the products will provide valuable initial templates for use of the Bayesian Models and overall ethical calibration system with a variety of current and future LAWS.

With respect to user interaction with the model, we were warned by several SMEs that completely autonomous ethical decision-making would not be currently accepted, though several acknowledged that it would be technically feasible within the decade. One SME cited Army Generals Rainey and Beagle, who both insist that the US will always require a human commander to make lethal decisions. This SME posited that "You cannot substitute computer systems for a smart guy." We concur with this vision and believe it strongly supports the use of an ethical model to speed the human decision-making process. Quick ethical situational awareness and decision aiding is especially crucial given the topical near-peer use cases recommended by our SMEs. These vignettes are persuasive for wargaming and for marketing our solution to military leadership.

Overall, the recommendations that came from the present trials will significantly improve the core product, which we hold to be sound. According to one SME, "This is the thought process that needs to be ingrained in commanders' heads."

Acknowledgements

We appreciate the help and guidance of our Technical Points of Contact Dr. Joseph Lyons, Principal Research Psychologist, Collaborative Interfaces and Teaming Branch, 711 Performance Wing, Air Force Research Laboratory (AFRL) and Ms. Brandi Kutter from the Air Force Material Command. We likewise appreciate the initial help provided by TPOC Mr. David Farrell, from the Air Force Material Command, before he left for another assignment. We also thank our consultant COL Jon Campbell (USA Ret) for his review of and contributions to our Bayesian Models and UAV scenarios based on his many years as a SOCOM operator and officer. And we acknowledge the valuable contributions of former military personnel to the test and evaluation of the ethical decision model and user interface. Finally, the scenarios used in the article are fictional and are not based on actual mission planning data that is restricted from public disclosure.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work reported here was performed under Air Force AFWERX SBIR Phase II Contract Number FA864922P1015 for Calibration of Ethical Trust in Lethal Autonomous Systems (CETLAWS).

Editorial disclaimer

The publishing of this article does not imply any promotion of Perceptronics Solutions as a company, or of its policies and products.

Notes on contributors

Spencer Kohn is the Director of Human Factors Research at Perceptronics Solutions; he is responsible for managing and contributing to research and development of human focused systems. His emphasis areas include calibrating trust in mixed human-automation teams through transparency, and maintaining situational awareness and balanced workload while using complex interfaces. Dr. Kohn has previously facilitated project development on topics including trauma medicine, combat jet piloting, and epidemiology; he has applied goal-directed task analyses to measure situational awareness in a virtual three-dimensional world, and designed and tested a Bayesian-based analytics and visualization system to supplement experts' evaluation of trainee military aviators. His current projects include the Intelligent Human Machine Interface (IHMI), an ongoing project with the US Army focused on assessing and adapting operator's trust, situational awareness, and workload in real-time based on non-obtrusive behavioral measures captured within the interface. Dr. Kohn received his Ph.D. in Psychology, Human Factors and Applied Cognition from George Mason University.

Marvin Cohen is an expert in cognitive systems and decision analysis. At Perceptronics Solutions, Dr. Cohen is a Senior Research Scientist and a lead investigator in projects centering on team performance and team cognition. He has been a principal investigator on several human – autonomy team performance efforts and helped to develop the Causal Model-based Measurement and Visualization System (CMVS) for Team Performance in Command and Control of Unmanned Systems for ONR. Dr. Cohen collaborated with Prof. Eduardo Salas on the development of Team Performance Framework for the Army Research Institute and developed team readiness measures for the Navy. Most recently his development of a cognitively efficient uncertainty visualization interface was transitioned to the Electronic Warfare Planning and Management Tool (EWPMT), a US Army Program of Record. Dr. Cohen has managed numerous projects involving the design, development, and testing of decision aids and research on human and normative decision processes. Dr. Cohen received his Ph.D. in Experimental Psychology from Harvard University and has over 150 publications in Journals, book chapters and conference proceedings.

Athena Johnson is a Principal Software Engineer at Perceptronics Solutions Inc. She specializes in designing and developing user interfaces across multiple domains that range from research and development to fielded mission critical systems. Her focus areas include analysis of interactions through conceptual interfaces, providing situational understanding through cyber warfare systems, ensuring resilience in mission critical systems and command and control through autonomous collaboration of unmanned systems. During her time at Perceptronics Solutions, Ms. Johnson has worked on projects including the Development and Run-time Environment for Aviation Mission-tasking and Mission Management (DREAMM), a system that provides the ability to develop complex mission compositions using an intuitive user interface and simplified semantics. She also participated in the Robotics Enhancement Program (REP), where she was able to successfully demonstrate unmanned command and control systems in live flight exercises. Ms. Johnson holds a B.S. in Computer Science from North Carolina State University and a M.S. in Computer Science from Hood College.

Mikhail Terman is a Data Scientist at Perceptronics Solutions, specializing in the application of generalized linear models, non-linear models, parametric and nonparametric tests, and time series analysis. At Perceptronics, Mr. Terman has analyzed and cleaned data in evaluating the use of new technologies for the U.S. Government, and has designed technical benchmarks for a number of AI based projects. Prior to joining Perceptronics, Mr. Terman collated and analyzed data on $250 million in loans to predict future losses that a private investment agency might incur, compared rating scales of foreign legal systems in order to predict loan repayment rates, tested proficiency of pre-existing models used by the department and found errors in active datasets and drafted a white paper on historical economic trends in Turkey and the performance of the country's financial sector to predict losses during the 2018 fiscal crisis. Mr. Terman received his B.A. in Psychology, *cum laude,* from Haverford College**,** and his M.S. in Statistical Science from George Mason University.

Gershon Weltman is an internationally respected expert in human factors and user interface design with a strong emphasis on training and simulation. Dr. Weltman's professional experience includes his long-term executive and technical management of our predecessor company Perceptronics, Inc., where his responsibilities as CEO included developing new business and directing the creative design, production and delivery of many innovative simulation and decision support systems. In addition to serving as principal investigator on a number of current Perceptronics Solutions projects, and providing scientific direction to others, Dr. Weltman is a Lecturer in the UCLA School of Engineering and Applied Science, where he teaches an undergraduate course on Engineering, Ethics and Society. Dr. Weltman was a six-year member of the U.S. Army Science Board, a select group that advises the Army on science and technology matters, and also served on the Defense Science Board Task Force on Training for Future Conflicts. He has published numerous scientific, technical, and strategic papers, and has presented lectures and briefings at government, business, and professional meetings. He holds a B.S., M.S., and Ph.D. in Man-Machine Systems from the UCLA School of Engineering and Applied Science.

Joseph Lyons is a Principal Research Psychologist within the 711 Human Performance Wing at Wright-Patterson Air Force Base, OH. Some of Dr. Lyons' research interests include human-machine trust, interpersonal trust, human factors, and influence. Dr. Lyons has worked for the Air Force Research Laboratory (AFRL) as a civilian researcher since 2005, and between 2011 and 2013 he served as the Program Officer at the Air Force Office of Scientific Research where he created a basic research portfolio to study both interpersonal and human-machine trust as well as social influence. Dr. Lyons has published in a variety of peer-reviewed journals. He is an AFRL Fellow, a Fellow of the American Psychological Association, and a Fellow of the Society for Military Psychologists. Dr. Lyons received his Ph.D. in Industrial / Organizational Psychology from Wright State University.

References

Aikins, Matthieu. 2021. "Times Investigation: In U.S. Drone Strike, Evidence Suggests No I S I S Bomb." *New York Times*, September 10. Accessed October 5, 2022. https://www.nytimes.com/2021/09/10/world/asia/us-air-strike-drone-kabul-afghanistan-isis.html.

Arkin, Ronald. 2009. *Governing Lethal Behavior in Autonomous Robots*. New York: CRC Press.

Arkin, Ronald. 2010. "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9 (4): 332–341. https://doi.org/10.1080/15027570.2010.536402.

Bahner, J. Elin., Ante-Dorothea Hüper, and Dietrich Manzey. 2008. "Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience." *International Journal of Human-Computer Studies* 66 (9): 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001.

Department of Defense. 2011. *Unmanned Systems Integrated Road Map for FY2011*-2036.

Department of Defense. 2023. *DoD Directive 3000.09 Autonomy in Weapon Systems, January 25,* 2023.

Dershowitz, Alan. 2022. [Personal communication]. July 3, 2022.

de Visser, Ewart, Marvin Cohen, Amos Freedy, and Rajiv Parasuraman. 2014. "A Design Methodology for Trust Cue Calibration in Cognitive Agents." In *Virtual, Augmented and Mixed Reality: Designing and Developing Augmented and Virtual Environments*, edited by Randall Shumaker and Stephanie Lackey, 251–262. Cham: Springer.

Devitt, Susannah Kate. 2021. "Normative Epistemology for Lethal Autonomous Weapons Systems." In *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, edited by Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, 237–258. Oxford: Oxford University Press.

Filkins, Dexter. 2021. "Did Making the Rules of War Better Make the World Worse?." *The New Yorker*, September 6. Accessed April 15, 2024. https://www.newyorker.com/magazine/2021/09/13/did-making-the-rules-of-war-better-make-the-world-worse.

Freedy, Amos, Ewart de Visser, Gershon Weltman, and Nicole Coeyman. 2007. "Measurement of Trust in Human–Robot Collaboration." In *International Symposium on Collaborative*

Technologies and Systems 2007, edited by William K. McQuay and Waleed W. Smari, 106–114. Orlando: IEEE Publications. https://doi.org/10.1109/CTS.2007.4621745.

Galliott, Jai. 2021. "Toward a Positive Statement of Ethical Principles for Military AI." In *Lethal Autonomous Weapons*, edited by Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, 121–136. Oxford: Oxford University Press.

Hodge, Nathan. 2010. "U.S. Military Releases Redacted Records on 2007 Apache Attack, Questions Linger." *Wired*, April 2. Accessed October 15, 2022. https://www.wired.com/2010/04/military-releases-report-on-2007-apache-attack-and-questions-linger/.

Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50.30392.

Malle, Bertram F., and Daniel Ullman. 2021. "A Multidimensional Conception and Measure of Human–Robot Trust." In *Trust in Human–Robot Interaction: Research and Applications*, edited by Chang S. Nam and Joseph B. Lyons, 3–25. Amsterdam: Elsevier.

Onnasch, L., C. D. Wickens, L. Huiyang, and D. Manzey. 2014. "Human Performance Consequences of Stages and Levels of Automation." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (3): 476–488. https://doi.org/10.1177/0018720813501549.

Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans – Part A* 30 (3): 286–297. https://doi.org/10.1109/3468.844354.

Ponniah, Kevin, and Lazara Marinkovic. 2019. "The Night the US Bombed a Chinese Embassy." *British Broadcasting Corporation*, May 7. Accessed October 5, 2022. https://www.bbc.com/news/world-europe-48134881.

Shanahan, Murray. 2016. "The Frame Problem." *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), edited by Edward N. Zalta. Assessed November 5, 2023. https://plato.stanford.edu/archives/spr2016/entries/frame-problem/.

Shortland, Neil D., Laurence J. Alison, and Joseph M. Moran. 2019. *Conflict: How Soldiers Make Impossible Decisions*. Oxford: Oxford University Press.

US Army. 1993. *FM 100-5 Operations*. Washington, DC: Headquarters, Department of the Army. Accessed November 5, 2023. https://www.bits.de/NRANEU/others/amd-us-archive/fm100-5%2893%29.pdf.