

Chapter 6

Ethical, Legal, and Societal Implications of the DARPA *in the Moment* Program



Daniel Trusilo, Lauren Diaz, and Ellie Tyler

6.1 Introduction

In 2022, the United States Defense Advanced Research Projects Agency, known as DARPA, initiated a multi-phase, basic research program called “In The Moment” (ITM) with the objective of enabling human-off-the-loop delegation to tactical algorithmic decision-making (ADM) systems in high-stakes domains where there is no agreed upon right answer, such as battlefield triage (DARPA 2024a). In line with DARPA’s mission, “To make pivotal investments in breakthrough technologies for national security,” ITM is supporting the development of ADMs that are trusted by humans to independently make human-off-the-loop decisions when human lives are at stake. Dr. Patrick Shafto, DARPA’s ITM Program Manager, explains, “[ITM] addresses, in the broadest sense, a key question about how we should think about humans and AIs working together in the future” (Voices from DARPA 2024). Specifically, the program is exploring if autonomous algorithmic systems that are aligned to human attributes will be more trusted by their human operators than unaligned systems. Shafto describes this work as a “bit more forward leaning than many DARPA programs. At some sense, we’re anticipating technology that will come soon, as opposed to trying to develop that technology right away...If we can come away from the program with a clear statement of what we mean by alignment and why it matters, then I think we will have made a great contribution to society broadly” (Voices from DARPA 2024). Since its inception, the ITM program has included an ethical, legal, and societal implications (ELSI) component, which is the

D. Trusilo (✉) · L. Diaz · E. Tyler
University of St. Gallen, St. Gallen, Switzerland

Applied Research Laboratory for Intelligence and Security (ARLIS), University of Maryland,
Rockville, MD, USA
e-mail: daniel.trusilo@unisg.ch

© The Author(s), under exclusive license to Springer Nature
Switzerland AG 2026

B. Koch, D. Winkler (eds.), *Artificial Intelligence Ethics in Military Medicine and Humanitarian Healthcare*, Military and Humanitarian Health Ethics,
https://doi.org/10.1007/978-3-032-11331-3_6

focus of this chapter. Shafto explains why considering ELSI is essential, stating, “The question of alignment and whether it matters for delegation is one that could have much broader implications than any specific technology that we could develop. And the social and legal regimes in which we think about that also could have extremely broad implications” (Voices from DARPA 2024).

Before discussing the ELSI aspects of ITM, in Sect. 2 we will give an overview of the ITM program including its underlying motivations, structure, and the metrics that are being used to measure success. Next, in Sect. 3, we will discuss the ELSI work being done as an integral part of the ITM program. This discussion will present the findings of a March 2024 event that was designed to share ITM with external experts; build a community of interest; and foster collaboration and debate on operational, as well as ELSI aspects of the program. Section 4 will then present an overview of challenges that have emerged and discuss future directions of this research. Finally, Section 5, will offer a brief conclusion.

6.2 ITM Overview

DARPA seeks to be disruptive. Classic examples of technology that has emerged from DARPA’s storied history include stealth and GPS capabilities (DARPA 2024b; Alexandrow 2008). Dr. Matt Turek, Deputy Director for the Information Innovation Office at DARPA, explains that in order to achieve breakthroughs, DARPA invests in research that falls along a continuum. On one end of the continuum, DARPA invests in research communities in a particular problem-space that’s relevant to the Department of Defense (DoD). This focus is intended to address perceived gaps in DoD’s technical capabilities. On the other end of the continuum, DARPA invests in building transformative capabilities that can be rapidly deployed to meet pressing needs of the warfighter (Center for Strategic and International Studies 2024).

The ITM program is a basic research program that is designed to support the research and development of algorithms that are trusted to autonomously make decisions in difficult domains, such as battlefield triage and disaster relief. ITM instantiates “trust” as “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al. 1995). The ITM program hypothesizes that humans will be more likely to trust an ADM that aligns with, and demonstrates human decision-making styles. ITM’s approach to developing such a capability is focused on the possibility of identifying human values and attributes that underline decision-making and encoding them to be represented in an ADM. Additionally, battlefield triage presents an objectively complex and highly consequential setting for researching humans’ willingness to delegate decision-making authority to an algorithmic system that is intended to operate without human input or oversight. Although military medicine is the primary use case, an alignment framework for a trusted human decision-maker and an autonomous algorithm can be applied to other domains.

Addressing the fundamental questions related to human decision-making, conditions for delegation, and human-AI alignment will inform future development and adoption of autonomous systems.

6.2.1 Motivations for ITM

ITM conceptualizes how humans and autonomous algorithms can work together in the future and is motivated by scenarios specifically related to triage and mass casualty events. Real-world events indicate that a human-off-the-loop ADM for triage may be needed by modern militaries to a) improve triage decision-making in scenarios in which human capacity is overwhelmed, and b) operate beyond current capabilities during peer and near-peer conflicts. A key question emerges when considering events that overwhelm hospitals and medical personnel. Dr. Patrick Shafto asks, “under what condition would a human delegate life or death decision to an AI system?” (Voices from DARPA 2024). For example, a 2017 Las Vegas concert shooting resulted in 215 gunshot wounds, straining the local emergency care facilities and challenging human-run triage operations (Menes et al. 2017). Could an ADM, developed to conduct triage in alignment with trusted human counterparts, help in such a situation? In addressing this question about alignment, Shafto asks, “Does it matter to humans that the AI system reflects our decision-making preferences, our morals, ethics, and so on? And would that increase the chances of us delegating in a particular situation?” (Voices from DARPA 2024). Perhaps, but such an ADM will have to be able to operate with limited data, making decisions that humans trust, when there is no clear right or wrong answer.

The challenge of developing an algorithmic system that can operate when there is no clear right or wrong answer is exemplified by three categories of conflicts that were evident in the Las Vegas triage operations. These three conflicts are: (1) rule conflicts, (2) policy conflicts, and (3) resource conflicts. First, the Las Vegas mass casualty event highlights the challenge of abiding by standard rules in a dynamic, high-consequence situation. Dr. Kevin Menes, the triage surgeon leading the effort, discussed the way in which he categorized incoming patients, reporting: “By textbook standards, some of these first arrivals should have been black tags, but I sent them to the red tag area anyway. I didn’t black tag a single one” (Menes et al. 2017). Menes’s statement means multiple patients should have been labeled as non-recoverable according to modern medical standards, but Menes, the human decision-maker, chose to prioritize emergency medical care for these patients to accelerate initial patient processing. This is an example of a conflict with existing rules, which an operational ADM must take into account.

Second, the Las Vegas mass casualty event resulted in a conflict with policy rules as Menes chose to delegate triage responsibilities to a nurse who had been assisting him, despite policy stating triage should be run by the most experienced doctor (Menes et al. 2017). This raises the question, could an aligned ADM serve as a

viable, even necessary, alternative to a human decision-maker should the situation require non-standard solutions, potentially conflicting with existing policy?

Third, during the Las Vegas mass casualty event, medical resources were strained. With twenty-eight damage control surgeries happening in the first six hours after the event and every single ventilator being used, the Las Vegas hospital network had exceeded its capacity (Menes et al. 2017). If we look beyond ITM and consider an operational ADM, we can hypothesize that an ideal system will account for a constantly changing resource environment, in which non-standard uses of resources may be required to maximize life-saving potential. With the Las Vegas mass-casualty event serving as a compelling example, the ITM program is driven to address high-consequence decision-making challenges related to standard rules and policy, as well as resource constraints, when there is no clear right answer, and alternative outcomes cannot possibly be known.

A second real-world case that exemplifies the motivation behind ITM is related to battlefield casualties. Namely, combat casualty medical care in the ongoing conflict in Ukraine indicates that human-off-the-loop automation may be required in future conflicts. Specifically, there is evidence that Russia incurred between 1000 and 1200 casualties per day during the peak of the Battle of Bakhmut (Axe 2023) resulting in potentially 66,000 to 88,000 deaths in the span of 9 months (The Economist 2024). The triage personnel of any military, no matter how well trained, will be overwhelmed given similar numbers. Further, there may be limited human capacity to perform triage operations in the event that forward operating medical personnel are wounded or killed. ITM is motivated to address this gap by exploring how to build trusted systems that can improve decision-making in such a scenario, aiding medical operations, and possibly reducing the number of deaths.

Underlying these motivations are a series of technical challenges that raise explicit and implicit questions related to ELSI. For example, when a human decision-maker is faced with multiple casualties, how do they decide, under extreme time constraints with limited data points, who to treat first? Such triage decisions can impact who lives and who dies, and there is no way to check the outcomes of a different choice after the fact. Do human decision-makers performing triage automatically act in compliance with the law of armed conflict; treating enemy combatants, civilians, and friendly force casualties the same? What if a mission critical element depends on keeping a certain individual alive? And how do an individual's decision-making attributes, values, and priorities influence their decision-making? These are just a few examples of the multitude of factors that make it challenging to concretely identify, measure, and algorithmically replicate a trusted human's decisions. However, for an ADM to be successful, identifying, measuring, and algorithmically replicating these factors is exactly what must be done.

6.2.2 ITM Program Structure

To develop an ADM that can be used for triage situations, and given the range of challenges highlighted above, ITM is structured with four Technical Areas (TA) of research: (TA1) Decision-maker characterization, (TA2) Human-aligned algorithmic decision-makers, (TA3) Evaluation and, (TA4) Policy and practice integration (See Fig. 6.1: ITM Technical Areas).

As the name suggests, TA1 researchers are focused on characterizing human decision-makers; specifically, the human attributes or values that may be relevant to triage decision-making. To that end, this requires TA1 researchers to create computational representations of decision-makers that summarize their attributes given a set of domain-specific circumstances. In other words, if a medical decision-maker encounters three patients with different injuries and physical characteristics, how can one quantitatively measure characteristics of the human decision-maker that define how they make their decisions about the three patients? To do this, decision-maker attributes will need to be identified and defined; the impact of situational information, domain knowledge, and other contextual elements on decision-maker attributes will need to be understood; and decision-maker responses to scenarios and decision-points will need to be represented in a computational framework.

TA2 researchers are focused on building human-aligned algorithms. To do this in a way that can be scientifically measured, TA2 researchers will need to build ADMs with quantifiable alignment to trusted human decision-maker attributes. Considering the example above, this means that an algorithm or set of algorithms will need to include variables that account for the characteristics of human decision-makers who

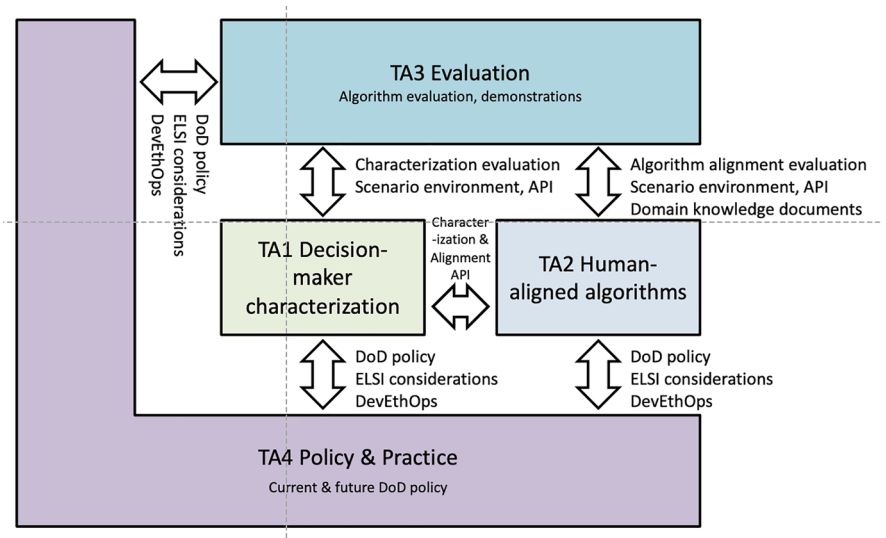


Fig. 6.1 ITM Technical Areas (DARPA 2024a)

are deciding how to triage the three patients in a way that is measurable. Such quantifiable alignment will allow for ADMs that can be adjusted to a group of trusted human decision-makers or even fine-tuned to a trusted individual decision-maker, such as the senior surgeon responsible for triage in a mass-casualty incident. Further, TA2 researchers must be able to integrate high-level goals and guidance such as DoD policy, Rules of Engagement, and commander's intent, into the ADMs.

TA3 researchers are responsible for evaluating the algorithms and demonstrating the technology. This means TA3 researchers must develop evaluation protocols, requirements, and metrics for triage decision-making in austere conditions and for a mass casualty event. Given the intent to discover if human decision-making attributes can be algorithmically encoded in an ADM, TA3 researchers must also design an environment that allows for data collection while accounting for psychological fidelity for human decision-makers. This will require a complicated array of objectives such as identifying a reference pool of trusted human decision-makers and ascertaining key knowledge in order to provide a control for ADM performance.

Ensuring that the systems being developed are in alignment with ethical and legal norms throughout the full system lifecycle is essential. Therefore, ITM includes an integrated policy and practice component via TA4. Specifically, TA4 is divided into two ELSI teams: one is responsible for internal ELSI components related to program design and ADM development and the other focuses on socializing the ITM program with a broad audience to build a community of practice and pathways to policy. This technical area is predicated on the understanding that long-term success of an ADM requires acceptance by the broader policy community as well as system end-users. Since there are limited policies governing ADMs to date, curating a multidisciplinary community of experts is a critical step in the conversation about potential policies that will be required for future technological capabilities. Additionally, incorporating policy and practice considerations from the start of the ITM program will aid developers across the other three technical areas by ensuring that ELSI concerns are addressed throughout the program lifecycle. This is especially relevant to the ethical, legal, and societal implications of human-aligned autonomous systems; to establish responsible and robust foundations for an ADM that will be used for human-off-the-loop decision-making when human lives are on the line and there is no clear right answer. In other words, ITM aims to develop a solution that accounts for policy and ethical concerns from the start so that it can be seamlessly deployed and responsibly incorporated into real-world operations.

6.2.3 ITM Metrics

ITM is using metrics to evaluate algorithmic performance that are guided by the program's objective to enable human-off-the-loop delegation to tactical ADMs in difficult DoD domains. The underlying approach taken by ITM to achieve this objective is to quantify the alignment of ADMs with trusted humans. This approach, and the related relationships between the TAs is visually shown in Fig. 6.2 below.

given alignment to a human expert (i.e. “you” or “your commander”) or a group of experts (i.e. a panel or board). ITM’s trust metric is founded in research that indicates medium correlation scores for trust measures range from 0.25 to 0.4, with strong results above 0.4 (Hancock et al. 2023). For a dry-run of its Phase I evaluation, ITM aims to achieve an intraclass correlation coefficient of the ITM alignment measure with a human trust scale of .3 using the Multi-Dimensional Measures of Trust v2 scale developed by Ullman and Malle (2023). This measurement addresses the need for ADM measures that are predictive of real-world, end-user trust. The concept of trust is inextricably linked with ELSI considerations. For instance, trust often develops through multiple interactions over time, in which one party acts in accordance with another party’s expectations. Further, trust requires vulnerability, in which the trusting party relies on the trusted party. Military medical experts, whom have consulted on ITM, have raised questions related to the cadence of reevaluating the relationship between a system-user and the system itself as well as the process for reestablishing user expectations when they are not met by the system. The experts emphasized that this reevaluation cadence will be important for the operationalization of an ADM in often ambiguous, high-consequence DoD environments.

The alignment measurement refers to the ability to tune an ADM to a subset of the attribute space. This measurement will be evaluated according to the ADM’s alignment to the largest cluster of human attributes. This measurement addresses the need for ADMs to be adaptable to key human decision-makers. Specifically, expert human decision-makers engaged in active combat on a high-priority operation will potentially make different triage decisions than an emergency medical professional in a civilian hospital. Decisions will be impacted by an individual’s values and ethical judgments as well as organizational policies and legal mandates. For an ADM to be operationally relevant, the ability to align to a specific human end-user or group will likely be required, be they a special operations medic deployed behind enemy lines or a team of trauma-surgeons facing a mass-casualty event in Las Vegas. That said, alignment has its own set of ELSI concerns: it could lead to group think, or, on the positive side, increased critical thinking and greater awareness of weaknesses. In an effort to test ITM’s hypothesis, it is equally important to consider broader questions, such as: Is alignment to a trusted human sufficient to establish trust in an ADM? How do we know if the alignment is “good enough”? Do we want to align an ADM to humans? And, if so, when is alignment appropriate?

Addressing these questions, as an integrated research area of ITM, encourages deeper consideration of the possible ethical, legal, and social implications that operationalizing an ADM for high-consequence decisions could pose.

6.3 A Deeper Look at ELSI

We have highlighted the integral nature of ELSI to the ITM program, but all of the issues raised require deeper discussion. At a fundamental level, ITM's approach to responsible AI is guided by, among others, the AI ethical principles of DoD's Chief Digital and AI Office (CDAO 2022) and those listed in the President's Executive Order on AI (The White House 2023). In fact, the CDAO's AI ethical principles and those elucidated in the 2023 Executive Order apply to all DoD AI capabilities, encompassing both combat and non-combat applications.

Clearly stating principles is a first step, but operationalizing ethical AI principles is challenging, requiring a practical approach based on understanding the real-world use case of any specific socio-technical system (Trusilo and Burri 2021). Therefore, ITM applies the CDAO AI ethical principles through an integrated technical area (TA4) that is focused on ethical, legal, and societal implications (ELSI) for the full program lifecycle including system development, testing, and evaluation. ITM, as a basic research program, does not include a deployment phase; however, TA4 research is informed by existing policies and practices, ensuring that ITM systems can be successfully operationalized, monitored, and revised as required. Additionally, TA4 researchers are socializing ELSI issues with a multidisciplinary audience of experts, hosting workshops designed to facilitate discussions among experts external to the ITM program on related policy, ethics, military medicine, triage, bioethics, AI alignment, and other relevant topics. These discussions play an essential role in establishing an interdisciplinary network that can be called upon to help identify and address gaps in existing policy.

To accomplish the required, full-lifecycle integration of ELSI, as described above, TA4 researchers are composed of two teams: (1) an internally facing advisory team led by the Institute for Defense Analysis (IDA), and (2) an outreach team led by the University of Maryland Applied Research Laboratory for Intelligence and Security (UMD ARLIS). The internally facing advisory team engages with ITM researchers from the other three technical areas in order to identify and advise on ELSI issues and provide expert guidance on relevant policies and practices.

The authors of this chapter are part of the outreach component of TA4, tasked with engaging with the DoD, academic communities, and other external stakeholders. In this role, we have organized two events designed to bring together those stakeholders and ITM researchers. We will now discuss some of the key points of discussion from the second ITM external outreach event.

6.3.1 External Outreach Event Overview

Long-term success of ITM depends in part on acceptance of the ADM by the policy community. Therefore, as part of the integrated ELSI effort, Phase I of the ITM program includes three external outreach events. These events are intended to

engage policy and practice subject matter experts from outside of ITM, with the goals of socializing the program; building a community of interest; and fostering collaboration and debate on operational, as well as ELSI considerations. In March 2024, the second outreach event (workshop) was held, which this section will discuss.

Development of the workshop was guided by three key questions: (1) How does policy influence decision-making in practice? (2) How were decision-makers able to adapt to other innovative techniques that disrupted traditional approaches? and (3) What can be learned about the potential to strategically incorporate automated tools in life-and-death decision-making for optimal outcomes? The objectives of the workshop were to provide members of the policy community, stakeholders, and ITM researchers a better understanding of: (a) human decision-making in situations where there is a high level of uncertainty and no right answers, (b) factors that may shape the willingness of a human to delegate decision-making authority to AI systems for life-or-death decisions where there is no right answer, and (c) how innovation can influence decision-making and relevant policy. Additional objectives specifically supported ITM researchers by elucidating: (d) future design elements that should be considered in ITM developmental processes, and (e) the technical information that is important for shaping future policy.

Subject matter experts (SMEs) in areas such as medical policy and decision-making, combat casualty care, and integrated military medical capabilities were invited to participate as panelists and speakers. The SMEs were affiliated with Army Futures Command's Medical Capability Development Integration Directorate (CDID), Walter Reed National Military Medical Center, Special Operations Medical Association, Harvard Medical School, DEU Joint Command Special Operations (German Army), Barts Health National Health Service (NHS) Trust and UK Defence Medical Service, Defence Science Organisation (DSO) National Laboratories in Singapore, Duke University, and the Deputy Assistant Secretary of Defense for Health Policy Readiness and Oversight. There were three panel discussions and one keynote presentation that offered perspectives on the benefits, risks, and challenges of applying autonomous decision-making technology to life-or-death situations.

6.3.2 Workshop Insights

Insights from the workshop suggest the adoption of an ADM depends on regulatory and cultural factors, and that a human will be more likely to delegate to a system if the system can be trained as a colleague in a skillset, such as tactical combat casualty care (TCCC), alongside a human operator. That said, workshop experts also highlighted that they do not always have the opportunity to train with the same human colleagues with whom they are required to operate. More specifically, the workshop highlighted: (1) variable levels of desire for human oversight, (2) the importance of addressing impacts on end-users such as the possibility of moral

injury, and (3) the challenge of rapidly establishing the trust of system operators and integrating new systems into existing military decision-making processes.

First, experts stressed the need for human oversight of an ADM but recognized the potential benefits of highly capable systems in high-stress, resource-constrained environments. During the workshop, international experts pointed out that military and medical training varies between countries and will impact the acceptance of ADM technologies. For example, cultural differences play a role in public perception and adoption of new technologies even though medical rules of engagement are standard across NATO and partner nations. Further complicating this challenge is the fact that cultural differences exist between services within any one country's military, which will impact any service branch's ability to successfully adopt disruptive technology. For example, U.S. Special Operations Forces (SOF) are more accepting of their highly trained medics making independent decisions that deviate from standard procedures.

Therefore, cultural differences must be accounted for when designing tactics, techniques, and procedures (TTPs) for an ADM use case. This is especially true because ITM is conceptualized as a human-off-the-loop system, designed to fill the space when human oversight is no longer possible because human decision-makers are overwhelmed or out of reach. In such an instance, will cultural differences matter or will the need to rapidly make high-consequence battlefield triage decisions in a highly variable environment force a non-conventional approach, such as delegation to an ADM? Ultimately, trust-building around ADM technology needs to be collaborative between government, public, and medical communities, taking into account the cultural contexts and maximizing touch points for system integration. Additionally, legal considerations for rapidly advancing technologies must be informed by existing law and detailed impact assessments of any ADM capabilities.

Second, the workshop highlighted the importance of addressing the possibility of moral injury of ADM end-users. In this context, the term moral injury was used broadly, referencing a range of definitions and implications in line with the wide-ranging research related to moral injury with regard to military populations reviewed by Richardson et al. (2020). Specifically, Richardson et al. state that the term *moral injury* broadly refers to the experience military service members often feel when “engaging in activities, witnessing acts, or immediate decision-making that may violate their moral codes and personal values. If unacknowledged, these factors can lead to injuries that can affect the physical, psychological, social, and spiritual health of military men and women” (2020).

An operational ADM presents the potential to lessen or exacerbate moral injury depending on the protocols applied to pre-, during-, and post-use. Additionally, consequences of such moral injury can vary from impacting a unit's overall mission readiness to creating new casualties. Therefore, choices that impact system use as well as the inherent Human-Computer-Interaction (HCI) of such use, must account for moral injury and the related design trade-offs.

At a basic level, it is widely accepted that an array of challenges exists when medical professionals choose to adopt AI technology that impacts treatment

provision (Rajpurkar et al. 2022). For example, if during a high-stress event, a medical professional does not understand a new technology's limitations and failure points, they can make a bad situation worse and hold themselves responsible when there is an undesirable outcome. To begin to address this concern, end-users can build trust in a system by training alongside it in a safe environment while also independently mastering the basics of combat triage. Additionally, workshop participants highlighted that the potential for moral injury and combat operational stress disorders could be lessened if end-users have an option to override an ADM without retribution. Therefore, embedding a non-retribution component, which allows users to override an ADM's decision, may reduce the risk of user moral injury. However, more research, along with thorough testing and evaluation protocols, and system certification procedures will be required to fully address the complexities of moral injury.

A third major point of discussion during the workshop was the concept of calibrated trust. Establishing trust typically requires progressive, incremental steps over time and is not instantaneous. Further, there are arguments that AI systems do not have the capacity to be trusted, but rather human interactions with AI systems is about reliance (Ryan 2020). But the mission space ITM is intended to operate in may not allow much time for system integration. When discussing this challenge, experts stated that ADMs could potentially become "technological colleagues" if trained alongside humans. Such concurrent training would demonstrate system reliability, credibility, care, respect, equity, and transparency. If this concurrent training was completed in multiple iterations, metaphorical "sets and repetitions" alongside a human counterpart, end-user reliance in the ADM would be engendered. Workshop participants stated that, ideally, an ADM would be considered a colleague rather than a tool controlled by a human. Such a system would be especially helpful for a population of individuals without a skillset, such as a military unit in which the medic is incapacitated or unavailable.

Additionally, several experts stated that while fully automated "colleagues" might be ideal, especially given a high-casualty environment, creating human-controlled tools that transparently present information is a vital initial step. The key is to make the space to carry out collaborative work with human end-users before there is no choice, when we must rely on fully automated systems because there is no other option.

During the workshop, it was also suggested that ADMs developed with physiology-based algorithms may better predict patient outcomes compared to ADMs with provider-oriented algorithms. This suggestion is based on the notion that predictors of human decision-making will have greater variability due to provider responses to environmental factors and emotions, such as fear, anger, or frustration. This raises a range of questions about for what an ADM designed for human-off-the-loop operation in high-stakes environments should be optimized. Ultimately, ITM is conducting basic research on alignment of an ADM with human decision-makers, based on human attributes, but future research could seek to identify performance differences between a provider-aligned ADM and a physiology-based ADM.

6.4 ELSI Challenges and Future Work

The workshop described in Sect. 3, as well as ongoing ITM ELSI work, highlights two overarching challenges of operationalizing ELSI considerations for disruptive technology designed for high-stakes DoD domains. These two overarching challenges, described in more detail in Sect. 6.4.1, include: (1) the need to address an array of conflicting policies and practices, and (2) the fact that ELSI considerations are highly nuanced. With these challenges in mind and guided by the specific needs of the ITM program, we are able to elucidate future ITM ELSI related work in Sect. 6.4.2.

6.4.1 *Overarching Challenges of Operationalizing ELSI Considerations*

First, any new system must be built on essential foundations, such as regulations, policies, and TTPs. For example, a tactical ADM designed for triage in austere conditions must be informed by existing DoD policies. However, ITM work must also inform the development of future policies and concepts of operations (CONOPs) to address areas in which existing policies, practices, and regulations are inadequate, irrelevant, or out-of-date for real-world use of a human-off-the-loop ADM. The only way policy gaps can be identified is if: (a) the system's use cases, capabilities, and limitations, are well understood, and (b) the existing body of relevant policies, TTPs, and regulations are well understood.

Identifying and addressing the body of material that captures relevant policies and practices must be done in a transparent and inclusive way. Additionally, all key stakeholders must be involved, including intended system operators, system developers, commanders of units that will deploy the systems, and partner nation military personnel. Because military commanders are legally responsible for the actions of the units they are commanding, CONOPs that embody a commander's assumptions and intent must be developed. Related, system certification models must be established, so that a commander knows a system is safe and reliable. For a program such as ITM, in which basic research on a disruptive technology is the goal, expertise that goes beyond current practice is required.

Second, early integration of ELSI considerations benefits ITM researchers' and developers' design considerations and the systems' ELSI-related capabilities, but addressing specific concerns in a nuanced way takes time and concerted effort. For example, as shown above, an ADM can perform to design specifications and still lead to moral injury for end-users. Such a possibility is a result of the inherent disconnect between developers, who require a specific expertise based on years of experience developing highly technical algorithmic systems, versus system end-users, who have a very different kind of expertise, requiring a comparable amount of time and effort to develop. Simply put, it is not enough to recognize that there is

potential for moral injury. Any program developing a system designed for high-stakes DoD missions, when there is no clear right answer, must be prepared to confront challenging questions and design tradeoffs. Not addressing the disconnect between system developers and end-users will have far-reaching impacts, such as negatively impacting an end-user's mission capability, even if a system operates exactly as designed. For this reason, ITM is committed to, for example, understanding the current degree of risk of moral injury for military medical personnel and how an ADM might lessen or intensify such a risk. Recognizing the challenge of addressing nuanced ELSI considerations is a fundamental reason why DARPA has chosen to integrate a policy and practice technical area for the ITM program.

6.4.2 *Future ELSI Work*

Additional ITM research effort must be put into defining gaps, uncertainties, and disconnects related to ELSI considerations with the objective of clarifying future AI policy development. For example, specific requirements for ADM user acceptance must be clarified and codified in related policy. Such work will make it clear what level of alignment is required, which will have tertiary impacts on training procedures, the specificity of user-data required to align a system, and the lifespan of a system-user relationship. Perhaps, research will indicate that alignment based on individual user feedback is highly correlated with user-trust, in which case it will be critical to know the minimum amount of time and energy required to provide a specific level of individual user feedback for any particular system. Alternatively, research may indicate that alignment to individual users is achievable and is correlated with higher levels of user trust, but that such alignment leads to poor medical outcomes, impacting overall system adoption and commanders' willingness to deploy such systems. Identifying gaps related to ELSI is crucial for clarifying the boundaries of ambiguous areas that will inform future policy development that applies to the full lifecycle of autonomous systems.

The example above raises another area that requires additional investigation. Namely, additional research is required to determine if human-AI alignment is a reliable and safe metric. In other words, how can ITM developers know that alignment to humans is sufficient? Should a system designed to make high-stakes decisions, that involve questions of life or death, be aligned to humans? And if so, which humans? Quantifying attributes related to human decision-making will require the consideration of human attributes that are likely to be unstable. Therefore, research must be done to determine if human decision-making attributes can be generalizable across high-stakes environments or if such attributes are domain-specific. Putting this into the context of ITM, we must ask, are the human decision-making attributes that are being used by ITM TA1 researchers sufficient? Or, does such alignment-related work require additional or other attributes? Military triage leaders with whom we have engaged have emphasized that in order to trust an ADM, evidentiary data indicating the system can make better decisions than a human is

required. This is anecdotal evidence that the bar for an algorithmic system is higher than that for humans. Regardless of the validity of this statement, collecting useful data about system performance that is sufficient to support end-user buy-in when there is no “correct” answer, will be challenging.

Other use cases of an ITM ADM must be explored in greater detail as the ELSI challenges will likely vary depending on the context. For example, it will be important to consider the use of an ITM ADM in high-stakes environments that are more permissive than a near-peer conflict (e.g., natural disaster). Lessons learned from the war in Ukraine, as well as both domestic and international disaster response operations, will be leveraged to inform such research. Exploring other high-stakes domains outside of battlefield triage will contribute to the research of human decision-making and willingness to delegate and support the development of a tool that is useful in a range of dynamic, austere settings, in which resources are limited and time is of the essence.

Other future ITM-related work can explore ELSI considerations related to the personalization of an ADM to individual users by integrating user feedback. Will such personalization impact a user’s willingness to delegate to an ADM system? And, would the integration of individual user feedback impact the medical outcomes? Such questions introduce the possibility of trade-offs between increased user trust in a system, which may increase its actual use, versus a reduction in ideal outcomes based on purely physiological measures that lead to ADM outputs that do not align to an individual end-user. These questions also raise an overarching consideration related to the way in which an ADM fits into existing workflows and TTPs, which will impact system adoption and, later, potential disuse. In other words, further exploration of ELSI challenges related to human-aligned autonomous systems and the impact of alignment on human trust to delegate decision-making authority to an ADM in high-stakes domains, is required.

6.5 Conclusion

This chapter provided an introduction to DARPA’s “In The Moment” (ITM) basic research program, including the motivation behind the program, how the program’s interdependent technical areas are structured, and the metrics that will be used to test the program’s hypothesis that ADMs encoded with human decision-maker attributes and values will be more likely adopted by humans. With this background, we then presented details of ITM Technical Area (TA) 4: Policy and Practice, which includes the program’s Ethical, Legal, and Societal Implications (ELSI) efforts. Though there are many outstanding questions, it is clear that integrating an ELSI component in basic research programs designed to develop revolutionary capabilities aligns with DARPA’s comprehensive and innovative approach to pushing the limits of emerging technology in a responsible way. To that end, key points of discussion raised during ITM’s March 2024 workshop included balancing human oversight with critical system needs, the importance of addressing impacts on

end-users, and the challenge of establishing “trust” in an algorithmic system designed to independently operate in high-stakes environments when there is no clear right answer. By highlighting both the range of outstanding policy and practice questions and the work being done to address them, this chapter shows the importance of DARPA’s emphasis on full integration of ELSI research. This research forces us to ask, how do we successfully and responsibly disrupt current military medical practices when time is of the essence? Ultimately, if human-off-the-loop systems become essential and the capability exists to build and deploy them, an approach that incorporates ELSI is not just best practice but essential to success.

Acknowledgments The work described in this chapter was only possible through the collaboration of dozens of individuals across multiple organizations. In particular, the authors wish to acknowledge the DARPA ITM program team, the members of other technical teams involved in the ITM program who actively participated in ELSI activities, and the many experts that provided candid and valuable insights at the ITM external events. The authors also deeply appreciate the contributions of their ITM ELSI teammates including David Danks, Brian Pierce, Harvey Rishikof, Rick Thomas, and Stephen Xenakis. Any errors or omissions are solely the responsibility of the authors.

References

- Alexandrow, C. 2008. *The Story of GPS*. [https://www.darpa.mil/attachments/\(2010\)%20Global%20Nav%20-%20About%20Us%20-%20History%20-%20Resources%20-%2050th%20-%20GPS%20\(Approved\).pdf](https://www.darpa.mil/attachments/(2010)%20Global%20Nav%20-%20About%20Us%20-%20History%20-%20Resources%20-%2050th%20-%20GPS%20(Approved).pdf). Accessed 15 May 2024.
- Axe, D. 2023. Did the Ukrainian Army Kill 1,100 Russians In A Single Day? It’s Certainly Possible. *Forbes*. <https://www.forbes.com/sites/davidaxe/2023/03/14/did-the-ukrainian-army-kill-1100-russians-in-a-single-day-its-certainly-possible/>. Accessed 1 June 2024.
- Center for Strategic & International Studies. 2024. *The DARPA Perspective on AI and Autonomy at the DOD*. Center for Strategic & International Studies Wadhvani Center for AI and Advanced Technologies. <https://www.youtube.com/watch?v=1xSw835-rig>. Accessed 20 October 2024.
- Chief Digital and Artificial Intelligence Office (CDAO). 2022. *Responsible Artificial Intelligence (RAI): Transforming the Department of Defense Through AI*. https://www.ai.mil/docs/CDAO_SLICKSHEET_RAIv2_07_25_22.pdf. Accessed 15 May 2024.
- DARPA. 2024a. *In the Moment (ITM)*. <https://www.darpa.mil/program/in-the-moment>. Accessed 1 June 2024.
- DARPA. 2024b. *DARPA’s Stealth Revolution*. <https://www.darpa.mil/about-us/timeline/darpas-stealth-revolution>. Accessed 15 May 2024.
- Goddard, K., A. Roudsari, and J. C. Wyatt. 2012. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association*. 19 (1): 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>.
- Hancock, P. A., T. T. Kessler, A. D. Kaplan, K. Stowers, J. C. Brill, D. R. Billings, K. E. Schaefer, and J. L. Szalma. 2023. How and Why Humans Trust: A Meta-Analysis and Elaborated Model. *Frontiers in Psychology*. 14:1081086. <https://doi.org/10.3389/fpsyg.2023.1081086>.
- Martin, A., C. L. Davidson, A. Panik, C. Buckenmyer, P. Delpais, and M. Ortiz. 2014. An Examination of ESI Triage Scoring Accuracy in Relationship to ED Nursing Attitudes and Experience. *Journal of Emergency Nursing*. 40 (5): 461–468. <https://doi.org/10.1016/j.jen.2013.09.009>.

- Mayer, R. C., J. H. Davis, and F. D. Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review*. 20 (3): 709–734. <https://doi.org/10.2307/258792>.
- Mistry, B., S. Stewart De Ramirez, G. Kelen, P. S. K. Schmitz, K. S. Balhara, and S. Levin. 2018. Accuracy and Reliability of Emergency Department Triage Using the Emergency Severity Index: An International Multicenter Assessment. *Annals of Emergency Medicine*. 71 (5): 581–7.e3. <https://doi.org/10.1016/j.annemergmed.2017.09.036>.
- Menes, K., Tintinalli, J., and L. Plaster. 2017. *How One Las Vegas ED Saved Hundreds of Lives After the Worst Mass Shooting in U.S. History*. <https://epmonthly.com/article/not-heroes-wear-capes-one-las-vegas-ed-saved-hundreds-lives-worst-mass-shooting-u-s-history/> Accessed 2 June 2024.
- Rajpurkar, P., E. Chen, and O. Banerjee. 2022. AI in Health and Medicine. *Nature Medicine*. 28:31–38. <https://doi.org/10.1038/s41591-021-01614-0>.
- Richardson, N. M., A. L. Lamson, M. Smith, S. M. Eagan, A. M. Zvonkovic, and J. Jensen. 2020. Defining Moral Injury Among Military Populations: A Systematic Review. *Journal of Traumatic Stress*. 33:575–586. <https://doi.org/10.1002/jts.22553>.
- Ryan, M. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*. 26:2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>.
- Skitka, L. J., K. L. Mosier, and M. Burdick. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies* 51:991–1006. <https://doi.org/10.1006/ijhc.1999.0252>.
- Tam, H. L., S. F. Chung, and C. K. Lou. 2018. A Review of Triage Accuracy and Future Direction. *BMC Emergency Medicine* 1 (58): <https://doi.org/10.1186/s12873-018-0215-0>.
- The Economist. 2024. *How Many Russian Soldiers Have Died in Ukraine*. <https://www.economist.com/graphic-detail/2024/02/24/how-many-russian-soldiers-have-died-in-ukraine>. Accessed 1 June 2024.
- The White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>. Accessed 2 June 2024.
- Trusilo, D., and T. Burri. 2021. The Ethical Assessment of Autonomous Systems in Practice. *J* 4 (4): 749–763. <https://doi.org/10.3390/j4040051>.
- Ullman, D., and B.F. Malle. 2023. *Measuring Human-Robot Trust with the MDMT (Multi-Dimensional Measure of Trust)*. <https://doi.org/10.48550/arXiv.2311.14887>.
- Voices from DARPA. 2024. 83: *When Should Machines Decide?* DARPA. <https://www.youtube.com/watch?v=vMIXAcEYfeU>. Accessed 30 October 2024.

Daniel Trusilo works at the intersection of AI, international security, and public policy. He is also a member of the Ethical, Legal, and Societal Implications team for the DARPA *In the Moment* (ITM) program. Dr. Trusilo earned a bachelor's degree from the US Military Academy at West Point, a Master of Arts in Law and Diplomacy from the Fletcher School at Tufts University, and Masters and Doctorate degrees in International Affairs from the University of St. Gallen, Switzerland.

Lauren Diaz is a research lead at the Applied Research Laboratory for Intelligence and Security (ARLIS), a University-Affiliated Research Center (UARC) at the University of Maryland (UMD), College Park. At ARLIS, Lauren supports intelligent human-machine systems research in the artificial intelligence, autonomy, and augmentation mission area. She is a policy and practice principal investigator for the DARPA ITM program. Lauren holds a Master of Public Health in Health Promotion, Education, and Behavior from the University of South Carolina, Columbia. Previously, Lauren worked as a senior research analyst at the University of Maryland, Baltimore's Institute for Innovation and Implementation. Additionally, Lauren coordinated large-scale projects funded by the National Institutes of Health (NIH) at the University of South Carolina, Columbia.

Ellie Tyler is a researcher at the Applied Research Laboratory for Intelligence and Security (ARLIS), at the University of Maryland. Joining ARLIS in 2023, she focused on the assessment of emerging technologies and their potential public health security implications. Ellie holds a bachelor's degree from Virginia Tech, where she studied Clinical Neuroscience alongside National Security and Foreign Affairs.