INSTITUTE FOR DEFENSE ANALYSES

# Ethical Considerations for the Use of Machine Learning in Military Personnel Management

## How should analysts behave? Philosophical foundations and action framework

## WEAI 2021

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley

June 2021
Approved for public release;
distribution is unlimited.
IDA Paper NS P-22652
Log: H 21-000159

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

**About This Publication**
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0045-14-D-0001, project BE-6-4311, "Assisting implementation of and analysis within DOD's Enterprise Data to Decisions Information Environment," for the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

**For More Information:**
Dr. Julie A. Lockwood, Project Leader
jlockwood@ida.org, 703-578-2858

ADM John C. Harvey, Jr., USN (ret) Director, SFRD
jharvey@ida.org, 703-575-4530

4850 Mark Center Drive
Alexandria, Virginia 22311-1882 • (703) 845-2000

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper NS P-22652

# Ethical Considerations for the Use of Machine Learning in Military Personnel Management

## How should analysts behave? Philosophical foundations and action framework

## WEAI 2021

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley

This page is intentionally blank.

# Executive Summary

Sound decision-making relies on the correct interpretation of relevant information. In principle, more information allows for better decisions. But information can be so vast and complex that no individual can synthesize it without algorithmic assistance.

Machine learning can help to distill oceans of information into something simpler and more directly relevant to decisions. At its core, machine learning consists of algorithms that are tailored to predict or classify a given outcome, such as predicting the probability that an individual will be able to successfully complete a training program based on a resume of existing skills, or classifying which position may be the most meaningful match for a new recruit. These algorithms are designed to identify and focus in on salient patterns within the data, ignoring information that is less relevant to predicting the desired outcome. This winnowing process enables such algorithms to ingest and identify intricate patterns in expansive quantities of data, resulting in much more accurate and detailed predictions than otherwise feasible. When used appropriately, such algorithmically synthesized information can empower human decision-makers to make faster, more consistent decisions. We refer to this as algorithmically assisted decision-making.

This Institute for Defense Analyses research seeks to clarify the foreseeable legal, moral, and ethical risks of machine learning and artificial intelligence in providing information used in personnel management processes, and to consider what can be done to mitigate these risks. The primary focus is on the military setting, but the underlying lessons apply to algorithmically assisted decision-making broadly.

Two levels of practical challenges confront researchers who support decisions with analyses:

- Analytical: How do we implement analyses that reflect the ethics and values of the organization and society?
- Communicative: How do we communicate with decision-makers about analyses to minimize inappropriate use?

This presentation addresses issues in analytical implementation.

Both algorithmically assisted decision-making and autonomous systems offer a growing number of promising applications to facilitate personnel management. Within the military context, the foundational personnel management objective is to acquire, develop, and retain personnel with the needed breadth and depth of skills, experience, and capabilities. Policies governing recruiting, occupational assignment, training, retention, promotion, force mix, command climate, family support, and other issues support the overarching goal. These policies can be shaped and adjusted

to enhance the well-being of the force. Information about attrition risks, personnel quality, recruiting effectiveness, and unit cohesion is vital to effectively shaping these policies. Information on many of these attributes exists as untapped potential within the vast personnel databases of the Department of Defense (DoD). Tapping and harnessing this potential requires synthesis tools, which machine learning techniques can help to provide.

Unfortunately, information can be misused. Misuses may be accidental, such as when a decision-maker predicates a decision on a misinterpretation. Other misuses, even if unintended, may directly violate law, morality, or ethical principles. For example, unlawful discriminatory outcomes are possible when decisions are made based on information that incorporates different patterns observed across race or gender lines. Likewise, privacy violations may occur through using, disclosing, or even inferring protected information. Information can also be misused when it is ignored. In some cases, information may not actually be misused, but those affected may perceive otherwise, eroding trust in institutions.

There are several reasons that machine learning models may aggravate the risk that information will be misused. First, the limitations of machine learning models may be ignored. Specifically, machine learning models develop predictions of historical outcomes based on historical data. If systematic errors or prejudices generated historical outcomes, then decisions based on machine learning models that do not account for those errors may perpetuate these problems. Machine learning models may likewise perform poorly in contexts that are sufficiently different from those represented by the underlying historical data.

Second, machine learning models can incorporate much more information than simpler statistical or heuristic models. This increases the risk of accidentally using data elements in an improper manner.

The third reason is that machine learning models may aggravate the risk of misuse of information. The internal workings of machine learning models can be difficult to understand, potentially making it more difficult to identify problems and reducing the trust of those affected.

A fourth reason is that although machine learning models can be highly complex, they also provide an explicit link between inputs and outputs. That explicit link can add transparency and traceability to processes that previously lacked such a link. While transparency is often viewed in a positive light, it opens new challenges. For instance, it may not be possible to simultaneously satisfy multiple ethical prerogatives. The transparency makes the failure to satisfy all ethical prerogatives more noticeable.

Finally, the precision of machine learning models can amplify damage from mistakes. Highly accurate machine learning-generated predictions can reveal things that perhaps should not be revealed, such as mental health conditions or pregnancy. Similarly, because machine learning can incorporate a broad scale of data representing large populations, mistakes can have far reaching consequences.

There is an emerging consensus on what society wants in machine learning and its applications, and while DOD has established its own ethical principles, these objectives often conflict or are impossible to objectively implement. How should we proceed when values conflict? Finding our way requires philosophical introspection. We set the stage with an overview of normative ethics, which is the branch of philosophy that describes and studies theories of moral behavior. We then explore ethical lessons and applications within the machine learning context. As new and unforeseen moral and ethical questions arise, researchers should assess these questions from foundational ethical principles. Only by examining foundational ethical principles and intentionally grappling with the ensuing contradictions can analysts provide consistent, reliable, and transparent analyses to decision-makers.

This page is intentionally blank.

# Ethical Considerations for the Use of Machine Learning in Military Personnel Management

## How should analysts behave?

## Philosophical foundations and action framework

## WEAI 2021

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley

6 May 2021

## Motivating examples:

*A military service uses an algorithm to synthesize officers' personal, performance, and training information into scores for each officer relative to specific opportunities.*

*An algorithm suggests service members for a retention bonus based on expected attrition date, career features, and performance history. Personal and family attributes influence the expected attrition date.*

# Applications of Machine Learning/Artificial Intelligence are promising, expansive, and often controversial

Stakeholders and analysts have many opinions on the morality of using ML/AI in human-centric applications

**Today's discussion is about the use of analytics to support human decision-makers**

# Counter: We have conducted complex, data-heavy analyses that affect people's lives for many years… What is different with ML/AI analyses?

Aspects of ML/AI analyses change how decision-makers experience, interpret, and act on our findings…

Higher fidelity results change what decisions are possible

Perception of increased accuracy makes action more likely

…and also increase the risk of problems from this action

Hype and plug-and-play toolkits from software firms—and sometimes from us—can dangerously over-represent results

Large volumes of analyses provide many opportunities for error

Large scale of analyses means large scale for potential problems

Public attention increases costs of real and perceived missteps

# Two levels of practical challenges confront researchers who support decisions with analyses

**Analytical:** How do we implement analyses that reflect the ethics and values of the organization and society?

**Communicative:** How do we communicate with decision-makers about analyses to minimize inappropriate use?

This presentation—and the associated paper—address issues in analytical implementation.

# Potential sources of ethical challenges in ML and AI

## Not all these items are bad or novel; all are challenging

**Inconclusive evidence**: Risk that algorithms output is incorrect

**Inscrutable evidence**: Source, scope, and quality of data used, and how data translate into results, may be unknown

**Misguided evidence**: Algorithms may use poor-quality or biased data, thus producing unreliable or biased results

**Unfair outcomes**: Results may lead to decisions with undesirable discriminatory effects, despite sound data and methods

**Transformative effects**: New and potentially unexpected insights can cause social, political, or perspective changes

**Traceability**: Identifying harms, the sources of harms, and whom to hold responsible may be difficult

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi, "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3, no. 2 (2016).

# Emerging consensus on what society wants in AI/ML…

Fairness and justice

Transparency

Interpretability and explanability

Accountability

Privacy

**Top five imperatives compiled in meta-analysis of:**
Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, et al., “The AI Index 2019 Annual Report,” AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, December 2019, https://hai.stanford.edu/research/ai-index-2019.

Thilo Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines* 30 (2020): 99–120; see p. 112.

Anna Jobin, Marcello Ienca, and Effy Vayena, “The Global Landscape of AI Ethics Guidelines,” *Nature Machine Intelligence* 1, no. 9 (2019): 389–99.

# …and DOD has established its own ethical principles…

| | |
|---|---|
| **Responsible** | "DOD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities." |
| **Equitable** | "The Department will take deliberate steps to minimize unintended bias in AI capabilities." |
| **Traceable** | "The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation." |
| **Reliable** | "The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles." |
| **Governable** | "The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior." |

# …but these objectives often conflict, or are impossible to objectively implement

*A military service uses an algorithm to synthesize officers' personal, performance, and training information into scores for each officer relative to specific opportunities.*

Officer Jane Doe wants to know her score and how it was calculated. Can she be provided complete transparency without compromising others' privacy?

How much transparency is appropriate? How does current transparency compare with that of legacy processes?

What aspects should be transparent:
The algorithm? The fitted model? The data?

Rhetorical question: How can the services leverage DOPMA* relief authorities without individual-level metrics?

* DOPMA stands for Defense Officer Personnel Management Act

# …but these objectives often conflict, or are impossible to objectively implement

*An algorithm suggests service members for a retention bonus based on expected attrition date, career features, and performance history. Personal and family attributes influence the expected attrition date.*

Suppose the suggested bonus pattern underrepresents some demographic groups. The data entering the model are confirmed as accurate.

Is the proposed bonus scheme *fair or unfair*?
Whose definition of fair should prevail?

Are some types of data off limits?
Are there some analyses that should not be conducted?

# How should we proceed when values conflict?

## Finding our way requires philosophical introspection

Ethics is the philosophical study of values systems

Moral judgments are applications of ethical frameworks

To resolve conflicts in values and moral judgments, we must appeal to ethical philosophy

Consistent values result from choosing and applying a cohesive ethical framework

# People make legal and moral judgments based on often-implicit *ethical frameworks*

## Consequentialist Ethics

*An action’s consequences determine its morality*

Utilitarianism – Ethical Egoism – Ethical Altruism

## Deontological Ethics

*Compliance with rules determines an action’s morality*

Rights- and Duty-Based Theories – Kantian Theory – Contractarianism

## Virtue Ethics

*Character traits of virtue, moral wisdom, and fulfilment constitute a person’s morality*

Eudaimonic (the greatest good) – Agent-Based – Target-Centered

# How do these ethical schools convert theory into action for ML/AI analysts and communicators?

## Consequentialist Ethics

Train models to discern between good and bad consequences or to maximize social benefits

## Deontological Ethics

Translate ethical principles into model design requirements

Fairness and justice
Transparency
Interpretability and explanability
Accountability
Privacy

## Virtue Ethics

Teach those conducting the analyses moral wisdom

# Law reflects society's ethical objectives

## A case study on anti-discrimination law

Deontological ethics are very influential in American law

Founders enshrine rights-based Jeffersonian values
Declaration of Independence, Constitution, Bill of Rights

Civil Rights Act of 1964 prohibits discriminating on the basis of race, color, religion, sex, or national origin

Intentional and unintentional discrimination both matter

Employers must establish *minimum standards* for worker characteristics to obtain a *business necessity exemption*

**Minimum standards can be demonstrated statistically**

# Questions for developers to ask themselves when operationalizing the DOD's ethical AI framework

**Planning:** Is machine learning the right tool for the job?

**Data selection:** Are the data appropriate to use and appropriate for the job?

**Design:** What issues or concerns should the developers be aware of in designing the machine learning model?

**Implementation:** What processes facilitate responsible use of the machine learning model?

# How our adversaries answer these questions matters

Weight placed on well-being of individuals vs. society will change with security conditions

How adversaries behave in a repeated game depends on

- Their values
- How their values evolve under various conditions
- Their beliefs about our values
- Their beliefs about how our values evolve under various conditions
- And so on…it's a complex game

Are there consequences to limiting our use of these tools?

How quickly can we change our posture?

- It takes time to develop and mature these capabilities

# Some final thoughts

Research leaders need to give serious thought to what analyses are undertaken and how they are conducted

Umbrella of *research* gives us some cover to learn without necessarily operationalizing models

Requires setting expectations with sponsor

My experience suggests that communication is far more difficult than research execution

Some uncertainty will be resolved with legal action

Ultimately, we are all accountable to our own consciences

IDA

| | |
|---|---|
| **REPORT DOCUMENTATION PAGE** | *Form Approved*<br>*OMB No. 0704-0188* |

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| | | |
|---|---|---|
| **1. REPORT DATE (DD-MM-YY)**<br>xx-06-2021 | **2. REPORT TYPE**<br>Final | **3. DATES COVERED (From – To)** |

| | |
|---|---|
| **4. TITLE AND SUBTITLE**<br>*Ethical Considerations for the Use of Machine Learning in Military Personnel Management*<br>*How should analysts behave? Philosophical foundations and action framework*<br>*WEAI 2021* | **5a. CONTRACT NO.**<br>HQ0034-14-D-0001<br>**5b. GRANT NO.**<br>**5c. PROGRAM ELEMENT NO(S).** |
| **6. AUTHOR(S)**<br>Alan Gelder<br>Julie Lockwood<br>Cullen Roberts<br>Ashlie Williams<br>Kathleen Conley | **5d. PROJECT NO.**<br>**5e. TASK NO.**<br>DZ-6-4720<br>**5f. WORK UNIT NO.** |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**<br>Institute for Defense Analyses<br>4850 Mark Center Drive<br>Alexandria, VA 22311-1882 | **8. PERFORMING ORGANIZATION REPORT NO.**<br>IDA Paper NS P-22652<br>Log: H 21-000159 |
| **9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**<br>USD(P&R)<br>1400 Defense Pentagon<br>Arlington, VA 22202 | **10. SPONSOR'S / MONITOR'S ACRONYM(S)**<br>USD(P&R)<br>**11. SPONSOR'S / MONITOR'S REPORT NO(S).** |

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Concerns about legal, moral, and ethical risks of machine learning have recently come to the forefront in the press, academical literature, and policy discussions. Do hiring algorithms risk running afoul of anti-discrimination laws? Can autonomous vehicles be trusted to weigh the ethical trade-offs of potential life-or-death situations? Does facial recognition software violate individuals' privacy rights? How can these and other potential problems be avoided or mitigated? Within the Department of Defense (DOD), concerns such as these are increasingly relevant as decision-makers seek to apply machine learning for a wide range of purposes. Building on foundational principles in ethical philosophy, this Institute for Defense Analyses presentation summarizes key legal, moral, and ethical criteria applicable to machine learning and provides pragmatic considerations and recommendations for its use in the personnel management context.

**15. SUBJECT TERMS**

Machine learning, personnel management

| | | | | | |
|---|---|---|---|---|---|
| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NO. OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON**<br>Lernes Hebert |
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | **19b. TELEPHONE NUMBER (Include Area Code)** |
| U | U | U | U | 28 | (703) 571-0114 |

This page is intentionally blank.