INSTITUTE FOR DEFENSE ANALYSES

Legal, Moral, and Ethical Implications of Machine Learning

DATAWorks 2022

Alan B. Gelder

March 2022
Approved for public release;
distribution is unlimited.
IDA Document NS D-33018
Log: H 22-000104

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22301

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
The work was conducted by the Institute for Defense Analyses (IDA) under CRP C6608.

For More Information:
Dr. John W. Dennis III, Project Leader
jdennis@ida.org, 703-845-2166
ADM John C. Harvey, Jr., USN (ret) Director, SFRD
jharvey@ida.org, 703-575-4530

Rigorous Analysis | Trusted Expertise | Service to the Nation

Executive Summary

Machine learning (ML) algorithms can help to distill vast quantities of information to support decision making. However, ML also presents unique legal, moral, and ethical concerns – ranging from potential discrimination in personnel applications to misclassifying targets on the battlefield. Building on foundational principles in ethical philosophy, this Institute for Defense Analyses presentation summarizes key legal, moral, and ethical criteria applicable to ML and provides pragmatic considerations and recommendations.

Addressing core legal and ethical concerns requires linking the ultimate action that an analysis plans to support with each step of the analysis. For instance, what data are appropriate to consider? Are the algorithm and the subsequent analysis effective for supporting the action?

This presentation also highlights implementation steps to maintain responsible machine learning and artificial intelligence at both the organizational level and the level of individual workers and analysts.

This page is intentionally blank.

IDA

Legal, Moral, and Ethical Implications of Machine Learning

DATAWorks 2022

Alan Gelder

April 28, 2022

Institute for Defense Analyses

730 E. Glebe Rd. • Alexandria, VA 22305

Ethics and Machine Learning/Artificial Intelligence (ML/AI) are increasingly in the public dialogue

Artificial Intelligence Index

Over 3,000 ML/AI news articles printed in 2020 included terms such as:
human rights, responsibility, fairness, discrimination, transparency, accountability, explainability, privacy

AI.gov launched in May 2021 to advance the design, development, and responsible use of trustworthy AI

Executive Order 13960 (Dec 2020) requires the Federal Government to use ML/AI with "due respect for our Nation's values," consistent with laws for "privacy, civil rights, and civil liberties"

IDA Stanford University's *AI Index 2021 Report* (p. 131). White House Press Release, 5 May 2021 1

Moving toward responsible use of trustworthy ML/AI

What **types of issues** do we need to be aware of?

What are **traits** of responsible ML/AI?

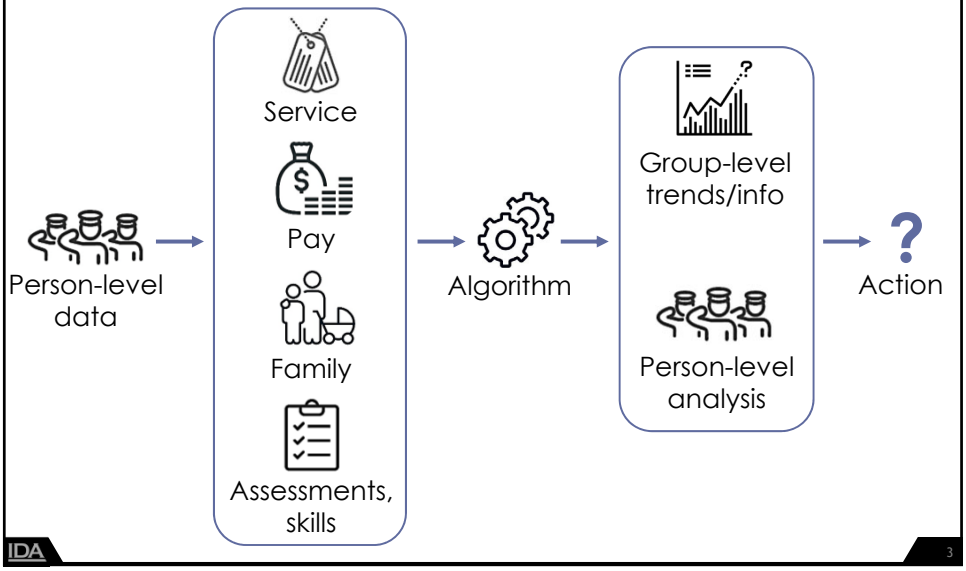
What are different **ethical approaches**?

How can we **implement** responsible ML/AI?

IDA 2

ML/AI ethics requires linking the ultimate action with the data, algorithm, and analysis informing the action

Example: Defense Personnel Management



IDA 3

Examples of actions for personnel applications of ML/AI

?
Action

Group-level action at arm's length
Budget plans informed by group-level trends

Group-level action targeting groups
Policy impacting group-level eligibility for a military career field (e.g., women in combat arms)

Person-level action for assessment processes
Scores used in career promotions/assignments

Person-level action for positive/negative screening
Enrollment in prevention program due to algorithmically determined higher propensity of risk

IDA 4

Is the analysis appropriate for informing the action?


Analysis

Hypothetical Example: An ML algorithm screens Special Forces applicants based on those who have served successfully over the last 20 years

How does the ban on women from combat arms until 2015 impact the analysis?

What are the legal requirements for avoiding potential discrimination issues?

What biases exist in the current system? Can ML do any better?

IDA 5

What data are appropriate to consider for the action?

Hypothetical Example: An algorithm synthesizes service members' career information into scores that impact selection for future opportunities

Data

Are there legal restrictions for how the data are used?

What is the boundary between career information to include and personal information to exclude?

How should the scores be built to properly exclude information deemed inappropriate?

Is the algorithm effective for supporting the action?

How much **transparency** is needed for each stakeholder?

Algorithm

How well does the model perform for individuals from different **demographic groups**?

How **interpretable** do the results need to be?

How is the model **maintained**?

How do you determine that the model is **effective** when testing procedures may be fragile?

Emerging consensus on what society wants in ML/AI

Top Traits from Meta-Analysis of AI Ethics Guidelines	
Fairness and Justice	Reduce or mitigate undesirable biases
Transparency	Openness of full ML/AI pipeline
Interpretability	Humans can understand relationship between inputs and outputs
Accountability	Responsive and defined liability
Privacy	Sensitivity in data collection and use

Top five traits 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. [59 documents reviewed]

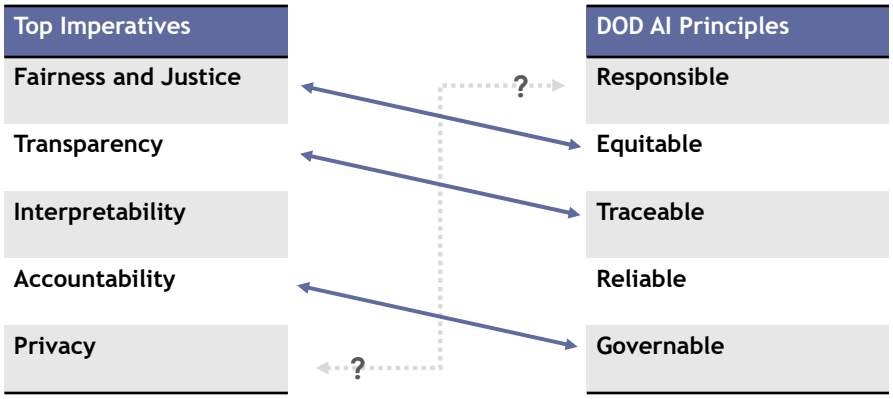
Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines* 30 (2020): 99–120; see p. 112. [22 documents reviewed]

Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389–99. [84 documents reviewed]

The 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."

Partial overlap between common AI ethical principles and the DOD's AI ethical principles



"Responsible" is a broad catch-all that can make it difficult to assess

There is a lot of agreement on the need for ML/AI ethics, yet the actual implementation is a challenge

Organizations issue high-level ethics principles, but without much clarity on what they mean in practice

Academics approach ethical questions in the abstract

On-the-ground workers understand institutional context, but often lack institutional support and academic rigor

 Reid Blackman, "A Practical Guide to Building Ethical AI," *Harvard Business Review*, 15 Oct 2020.

There are also different ethical approaches

Consequentialist Ethics

Consequences determine morality

Utilitarianism, Ethical egotism, Ethical altruism

Deontological Ethics

Compliance with rules determines morality

Rights & duty-based, Kantian, Contractarianism

Virtue Ethics

Moral character and virtues drive morality

Eudaimonic, Agent-based, Target-centered

And ethical approaches can blend and overlap

Example: Executive Order 13960

Federal agencies shall use AI …

When "the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed"

"In a manner…consistent with the Constitution and all other applicable laws and policies"

So how can we implement responsible ML/AI?

Organizational level

Articulate goals for meeting AI principles

What does winning and losing look like?

Invest in meeting AI principles

Build **responsive** ethical review into internal processes

→ *Agile, Timely, Authoritative*

Cultivate a virtuous workforce

Cannot regulate everything…

Teach and instill a virtuous ethos and how it applies to AI

So how can we implement responsible ML/AI?

Worker level

Planning: Is ML/AI the right tool for the job?

Status quo, alternative approaches, desired goal

Data Selection: Are the data appropriate for the job?

Accuracy, privacy, bias, discrimination, legal use

Design: What should developers be aware of in designing the model?

Stakeholder concerns, diversity, testing, transparency

Implementation: Are there processes to enable the responsible use of the model?

Training, documentation, monitoring, safeguards

Alan Gelder
agelder@ida.org

This page is intentionally blank.