INSTITUTE FOR DEFENSE ANALYSES

Legal, Moral, and Ethical Implications of Machine Learning for Personnel Management

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley
Rachel Augustine

April 2023*
Approved for public release; distribution unlimited.
IDA Paper P-33087
Log: H 22-000189

INSTITUTE FOR DEFENSE ANALYSES
730 E. Glebe Rd
Alexandria, VA 22305

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
This work was conducted by the Institute for Defense Analyses under contract HQ0034-19-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.

Acknowledgments
We would like to thank William Doane, Peter Levine, Peter Picucci, and Brian Williams for their careful review and helpful comments. This work also benefited from researcher feedback at the Western Economic Association International (WEAI) 96th Annual Conference, including Colin Doyle who served as a discussant. This acknowledgement does not indicate an endorsement by these reviewers.

For More Information:
Dr. Alan B. Gelder, Project Leader
agelder@ida.org, 703-845-6879
ADM John C. Harvey, Jr., USN (ret) Director, SFRD
jharvey@ida.org, 703-575-4530

***This research was conducted from March 2020 to May 2022.**

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-33087

Legal, Moral, and Ethical Implications of Machine Learning for Personnel Management

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley
Rachel Augustine

This page is intentionally blank.

Executive Summary

***This research was conducted from March 2020 to May 2022.**

Sound decision making relies on the correct interpretation of relevant information. For personnel management, the information may be subtle and interpersonal — the grip of a handshake, the sincerity of a smile. Or it may be rigorously codified and organized in databases, such as performance evaluations and service records. In principle, higher quality information allows for better decisions. Information quality is not synonymous with information quantity.

However, to the extent that quality information can be gleaned from the vast quantities of information that exist in a data-rich world, machine learning algorithms can help to distill this information into something simpler and more relevant to decisions. To meet the quality threshold for decision making, this information needs to meet appropriate standards of validity, legality, ethicality, or other relevant metrics.

Relative to more traditional analytic processes, machine learning algorithms (and the models incorporating them) may aggravate the risk that information will be misused. Machine learning algorithms automate the process of "learning" by using data to gradually improve performance relative to an objective. However, the limitations of these algorithms may be ignored. For example, machine learning models are frequently "trained" by using characteristics of past observations to predict outcomes that occurred in the past. Patterns observed in the past are then used to predict future outcomes. If systematic errors or prejudices generated historical outcomes, then decisions based on machine learning models may perpetuate these problems.

Machine learning models may likewise perform poorly in contexts that are sufficiently different from those represented by the data used to train the model. For example, trying to use the patterns that exist in one population to predict results for an entirely different population, or using historical patterns to predict future outcomes when the underlying historical conditions that generated the patterns are not representative of likely future conditions. Another challenge of models using machine learning is that, by incorporating much more information than more traditional methods, they can be highly complex and difficult to understand. This complexity can make it more difficult to identify problems and may also reduce trust in the models and the institutions using them.

Despite this complexity, machine learning models do provide an explicit link between inputs and outputs. Such an explicit link opens further challenges. For instance, because some ethical prerogatives cannot be satisfied simultaneously, this transparency can force individuals and organizations using these models to rank ethical prerogatives that they would prefer not to rank.

The precision and scalability of machine learning models can also amplify the potential for damage.

This paper attempts to clarify the foreseeable legal, moral, and ethical risks of machine learning — as well as what can be done to mitigate these risks — when applied to personnel management processes. Although the primary focus is on the military setting, the underlying lessons apply broadly to personnel management in a variety of contexts.

Our analysis begins with an overview of machine learning's strengths and weaknesses relative to more traditional statistical tools. We then present a handful of fictional vignettes to illustrate how machine learning could be implemented within military personnel management. These fictional vignettes address the following topics:

- *Using machine learning to produce summary metrics that are considered during the officer promotion process* — Since 2013, Department of Defense (DOD) policy has explicitly permitted promotion boards to "consider automated computer summaries of information in an eligible officer's official military personnel record" (DOD Instruction 1320.14, paragraph 3.2.c.(2)(a)). Given this policy guidance, as well as statutory limitations on the types of data available to a promotion board, are there appropriate uses of machine learning in the promotion process?
- *Using machine learning to facilitate recruiting for high-demand, low-density career fields, such as Special Forces* — Are there appropriate uses of machine learning to better target recruiting efforts to individuals who exhibit traits similar to those who successfully complete long, costly, and high-attrition training regimes?
- *Using machine learning to better budget and plan for training slots* — When a machine learning application relies on potentially sensitive personnel data, but the information is used for aggregate forecasts and planning, what ethical considerations apply?
- *Using machine learning for targeted retention interventions* — How might machine learning appropriately be used to enhance and augment targeted retention interventions?

Importantly, the technologies presented in these vignettes are not science fiction but could be implemented now or in the near future; thus, the pertinent questions are whether and how the technologies *should* be implemented. What guardrails are needed to avoid legal and ethical pitfalls, and what practices should be adopted to enable the responsible use and expansion of these technologies?

There is an emerging consensus on what society wants in machine learning and its applications. Fairness, transparency, interpretability, accountability, and privacy rank at or near the top of numerous ethical guidelines for machine learning and artificial intelligence. The DOD has likewise established its own ethical principles: responsible, equitable, traceable, reliable, and governable. In practice, however, these objectives often conflict. An inability to objectively measure whether a given principle has been satisfied may also cause some principles to be

prioritized over others. 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. The three major schools of thought — Consequentialism, Deontology, and Virtue Ethics — offer contrasting viewpoints on how to approach moral determinations, and certain aspects of each intersect with essential American principles and military values. It is these intersections that we find most applicable to the military personnel context.

To an extent, laws are a codification of society's ethical objectives. Given the recency of many machine learning advances, specific legal and regulatory requirements regarding appropriate uses of machine learning in personnel management processes are not fully delineated and are evolving. Personnel management decisions that are not lawful in the absence of machine learning are presumably not lawful in its presence. However, the presence of machine learning may invite a higher level of public and legal scrutiny to ensure that the law is indeed being followed.

Antidiscrimination law, in particular, presents some unique challenges that courts will likely need to address in the context of machine learning. Even without intent to discriminate, an employer could still have policies, procedures, or infrastructure in place that result in unequal outcomes across protected classes or other demographic groups of interest. However, outcomes do not always reflect discriminatory treatment. A job that routinely requires workers to carry loads of 100 pounds or more is likely going to employ more men than women. We provide a short examination of how antidiscrimination legal concepts such as *disparate impact*, *business necessity*, and *minimum standard* may relate to statistical thresholds used in machine learning contexts.

Implementing ethical and legal norms within machine learning models is complicated by the fact that machine learning models are highly dependent on the data they are training on. Human behavior and culture are complex, and data on individuals are likely to reflect the varied nuances of society — both the good and the bad. To the extent that the data reflect societal virtues and vices, a model that is trained on those data will likewise reflect and amplify these. Diagnosing and correcting undesirable patterns is left to human hands.

To a limited extent, machine learning models can impose compensating corrections that can help close the gap across demographic groups for some dimensions of inequality. Mathematically, not all dimensions can be closed simultaneously, so the user must determine a dimension to focus on. We provide a survey of machine learning methods that may be used to identify and (potentially) compensate for equality concerns. However, these corrections may only be legal for personnel management in limited circumstances.

A key consideration in introducing machine learning into personnel management practices is to compare current practices with how they might be conducted using machine learning. Executive Order 13960 (3 December 2020) directs Federal Government agencies to use machine learning and artificial intelligence when the "benefits of doing so significantly outweigh the risks, and the

risks can be assessed and managed." Can a machine learning application in a given decision-making process maintain or exceed the ethical status quo of current personnel policy? Can it move personnel policy closer to an ethical ideal? What is the potential harm if the machine learning application is misapplied, particularly in ways that could reduce the ethical status quo?

These questions likely have complex answers, necessitating a closer look at what is gained and what is lost. Much depends on the underlying design of the machine learning, such as having appropriate data inputs and transparent processes, effectively communicating how to use an algorithm's outputs to potential users, and incorporating measures for identifying and assessing risks into various stages of testing and operationalization. We propose a framework for considering issues like these throughout the full life-cycle of a machine learning project. This is summarized in the table on the following page.

We then return to the fictitious vignettes of machine learning applications for military personnel management to explore how the DOD's artificial intelligence (AI) Ethical Principles might be employed in practice. This discussion is not intended to provide an ultimate determination of whether these scenarios constitute legal and ethical use of machine learning, but rather to illustrate how stakeholders may approach similar deliberations.

Machine learning models cannot be developed, deployed, and simply forgotten. The predictive power of machine learning models trained on historical data can diminish over time. In adopting and using machine learning models in personnel management contexts, it may be appropriate to establish ad hoc or standing review bodies to provide oversight for how machine learning models are used. The ethical risks of using machine learning models in specific personnel management applications may not be immediately apparent. It may therefore be prudent to deploy the new process incrementally to provide learning opportunities while also reducing the scope for potential harm. Documentation may need to be developed both at a technical level and at a level accessible to users. Lay users should be able to understand what the model does and does not consider.

Ethical norms and standards for machine learning are evolving as society responds to and considers the risks and benefits of this technology in different applications. Ethical risks for some applications are real, but so are the ethical benefits for using this technology in many cases. Ethical codes and regulations should take these risks seriously, but in a way that balances development costs and the ethical consequences of *not* developing methods that may synthesize information better or more transparently. Tools to identify these risks and mitigate them are in flux. Approaches that overly fixate on particular mitigation methods may be counterproductive. Instead, we recommend a general ethical framework and general processes that can be adapted and applied on a case-by-case basis.

Life-Cycle Considerations for Machine Learning Projects

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

- What is the ultimate goal? If there are multiple goals, what are the trade-offs? Will machine learning provide appropriate information for achieving these?
- Would a different approach be more effective than a machine learning prediction?
- How would the risks and benefits of using machine learning differ from the status quo (including development risks, financial risks, legal and ethical risks, etc.)?

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

- Why were the data collected? How reliable are the data? Can their provenance be determined? Were there limits imposed on their use at the time of collection?
- How were the data prepared for analysis? What procedures are in place to ensure the data are maintained with the highest level of accuracy?
- Are there processes to periodically check for unintended bias within the data?
- Are there data elements or correlates of data elements that may be impermissible to consider for particular decisions? How will such data elements be dealt with?
- What safeguards are in place to respect privacy and protect the data?

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

- Have the developers consulted with subject matter experts, stakeholders, and groups that will be affected about potential legal, moral, or ethical issues that may arise?
- What dimensions of diversity should developers consider in designing the model?
- Is there end-to-end transparency in the machine learning project, from data collection and acquisition, to data preparation, to model coding, refinement, and testing?
- How will the system be tested and monitored? Is there a robust code review process? Is the pipeline fully reproducible? What tests will be done to monitor the adequacy of results (e.g., using a set of test cases that should be classified in a given way)?
- Is the pipeline sufficiently modular so that if a component is later found to be problematic, it can be readily swapped out?
- Are there corrective actions that should be taken to minimize differential outcomes across given populations? How will corrective actions be evaluated and tested?

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

- Do stakeholders and users understand the appropriate uses and limitations of the machine learning model? Is there effective documentation and training?
- What is the plan for monitoring and evaluating use of the model?
- Are safeguards in place to identify and intervene in case of unintended consequences?

This page is intentionally blank.

Contents

1. Introduction

Sound decision making relies on the correct interpretation of relevant information. For personnel management, the information may be subtle and interpersonal — the grip of a handshake, the sincerity of a smile. Or it may be rigorously codified and organized in databases, such as performance evaluations and service records. Information quality is not synonymous with information quantity.

However, to the extent that quality information can be gleaned from the vast quantities of information that exist in a data-rich world, machine learning algorithms can help to distill this information into something simpler and more relevant to decisions. To meet the quality threshold for decision making, this information needs to meet appropriate standards of validity, legality, ethicality, or other relevant metrics.

Machine learning algorithms automate the process of "learning" by using data to gradually improve performance relative to an objective. These algorithms are often tailored to predicting or classifying a given outcome, based on the connections between observed traits and associated outcomes observed previously. The nature of the outcomes can be vast and varied, such as predicting the probability that an individual will complete a training program, or classifying which position may be the most meaningful match for a new recruit.

Although prediction and classification algorithms are not new, machine learning algorithms amplify the scope, scale, and interactions within data that can be meaningfully used. These algorithms are designed to identify and focus on salient patterns within the data and ignore information that is less relevant to predicting the desired outcome. This winnowing process enables these algorithms to ingest and identify intricate patterns in expansive quantities of data, often resulting in much more accurate and detailed predictions than otherwise feasible. Such algorithmically synthesized information can empower human decision makers when used appropriately. However, machine learning algorithms do have limitations and risks that need to be addressed and carefully considered.

This paper attempts to clarify the foreseeable legal, moral, and ethical risks of machine learning — as well as what can be done to mitigate these risks — when applied to personnel management processes. The primary focus is on the military setting, but the underlying lessons apply broadly to personnel management in a variety of contexts.

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 is 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 are frequently "trained" by using characteristics of past observations to predict outcomes that occurred in the past. Patterns observed in the past are then used to predict future outcomes. 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 data used to train the model — such as trying to use the patterns that exist in one population to predict results for an entirely different population, or using historical patterns to predict future outcomes when the underlying historical conditions that generated the patterns are not representative of likely future conditions.

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: for example, inappropriately acting on a correlative relationship in the data as if it were a causal relationship.

Third, the internal workings of machine learning models can be difficult to understand, potentially making it more difficult to identify problems and reducing trust in the models and in the institutions using them.

Fourth, 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. Transparency is often viewed in a positive light, but it opens new challenges. For instance, it may not be possible to simultaneously satisfy multiple ethical prerogatives. The transparency makes the failure to satisfy multiple ethical prerogatives more noticeable, and it can force individuals and organizations using these models to rank ethical prerogatives that they would prefer not to rank.

Fifth, the precision of machine learning models can amplify the potential for damage. Machine learning-generated predictions can reveal things that data subjects may prefer remain private, 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.

A burgeoning literature discusses these and other risks of machine learning, as well as how to mitigate them. This paper draws on the literature while addressing its shortcomings. Shortcomings can range from an overreliance on buzzwords or vague terminology[1] to a failure to adequately relate cited risks or recommendations to any overarching ethical frameworks. This is especially problematic when ethical recommendations conflict and risks must be prioritized. Moreover, many of the legal and ethical concerns that apply to other applications do not apply in the same way to military personnel management.

We begin by providing a somewhat non-technical description of machine learning and some potential applications to military personnel management. This is important because anxieties about machine learning are sometimes borne from misunderstanding. Moreover, the risks of misusing or misinterpreting machine learning outputs are greater for those less familiar with its strengths and limitations. Chapter 2 ends with a series of potential use-case vignettes for how machine learning may be applied within concrete military personnel applications. These vignettes provide context for assessing the benefits and risks of machine learning along legal and ethical dimensions.

In Chapter 3, we explore various ethical frameworks from which to view machine learning applications, starting with a background of deontological, consequentialist, and virtue ethics. These ethical frameworks undergird applied ethics, even when the relationship is not explicit. How policymakers weigh ethical applications will depend on how they view these overarching ethical frameworks. This is especially important because, in many cases, it may be impossible to satisfy different ethical prerogatives simultaneously.

We describe laws and jurisprudence relevant for the application of machine learning to personnel policy in Chapter 4. This is an evolving area of law that is not always clear. We focus

---

[1] "Establishing a common language is essential for stakeholders to accurately be able to discuss issues. Since ML is an evolving field of computational science, misconceptions and myths are bound to emerge in the public dialogue with terms and phrases that may, or may not, be relevant to the practice of ML in lending. Buzzwords and vague or inconsistent definitions are often counter-productive for nuanced discussion of complex topics…" (*Machine Learning: Considerations for fairly and transparently expanding access to credit*. BLDS LLC, Discover Financial Services, and H20.AI., July 2020, 1st edition).

on antidiscrimination law and trace prominent cases that help to illustrate the broader legal environment and which may hint at the direction of new legislation and jurisprudence.

Having established the legal and ethical background, in Chapter 1 we illuminate practical ways to evaluate and mitigate legal and ethical risks presented by machine learning. This is important because many "ethical frameworks [for machine learning] cannot be clearly implemented in practice," which can create "the false sense that organizations have made their [machine learning] free of risk when its dangers are in fact rampant."[2] Some potential harms can be quantified, like disparate impact or proxy discrimination.[3] Tools can be applied to machine learning so as to reduce these harms. We discuss these tools, delineating how the ethical harms they are meant to mitigate relate to the overarching ethical framework. This discussion also incorporates mathematical definitions of fairness, which we use to help illustrate the incompatibility of different ethical goals.

The final two chapters outline practical guiding principles for operationalizing machine learning. These include institutional checks and processes which can help to ensure that legal and ethical risks are considered. Chapter 6 includes a checklist of life-cycle considerations for machine learning projects that can help guide the use of these tools from planning, to data selection, to design, to implementation. Chapter 6 also returns to the machine learning use-case vignettes from Chapter 2 to explore risks and other practical considerations in greater depth. We examine these vignettes within the context of the Ethical Principles for Artificial Intelligence that the DOD adopted in 2020.[4]

The DOD Ethical Principles are necessarily broad given the scope of DOD applications. Given their breadth, it is easy to lose focus on how the DOD Ethical Principles can be applied in practice. The examination of these vignettes illustrates actionable steps that may be taken on issues that will likely arise in the personnel management context. We conclude in Chapter 7 with additional recommendations on processes that can enable an institution to responsibly use machine learning capabilities.

---

2 Andrew Burt, "Ethical Frameworks for AI Aren't Enough," *Harvard Business Review*, November 9, 2020, https://hbr.org/2020/11/ethical-frameworks-for-ai-arent-enough.

3 Since the intent to discriminate (*disparate treatment*) can be difficult to prove in court, enforcement of the Civil Rights Act of 1964 (Pub. L. 88–352) is often based on showing that the outcome of a policy or practice disproportionately hinders a particular demographic group (*disparate impact*). *Proxy discrimination* is the use of a facially-neutral characteristic in a decision-making process, where the characteristic is correlated with a given demographic group, and its use masks open discrimination against that group.

4 See Department of Defense, "DOD Adopts Ethics Principles for Artificial Intelligence," DOD Newsroom, 24 February 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

2. Machine Learning for Personnel Management

The prominence of machine learning (ML) and artificial intelligence (AI) have grown rapidly in recent years. Although many of their core algorithms have existed for decades, only recently have computing capabilities reached the necessary level to effectively execute these algorithms on large enough datasets to achieve meaningful performance.

Vast and varied data are the fuel for ML and AI. Supervised learning algorithms — which account for a large portion of ML applications — require a substantial amount of data where the outcome of interest has been observed. The algorithms then use these data to identify patterns linking the inputs with the outcomes. This process is known as "training," where knowing both the inputs and the outcomes is akin to having training wheels on a bicycle.

Like the training wheels, information about outcomes will eventually be removed. The ultimate goal is to provide new data inputs and have the algorithm accurately predict the unobserved outcomes. The precise nature of the patterns linking the inputs with the outcomes can be highly complex. By "learning" these patterns, these algorithms become capable of predicting the outcomes. A central goal of ML is thus to synthesize data in order to make predictions.

ML is often viewed as a subset within the broader field of AI. For instance, the statutory definition of AI in Section 238 (g) of the National Defense Authorization Act (NDAA) for Fiscal Year 2019 (Pub. L. 115-232) refers to ML as one of many sets of techniques that constitutes AI. This definition includes five goals that an AI system may be designed to accomplish: (1) perform "tasks under varying and unpredictable circumstances without significant human oversight;" (2) solve "tasks requiring human-like perception;" (3) "think or act like a human;" (4) "approximate a cognitive task;" and (5) "act rationally."

These five goals can be summarized within two broad categories for how ML and AI can be used in policy making. The first is to synthesize or identify patterns in data (e.g., goals 2 and 4 in the NDAA definition). The second is to proxy for human-like action or choices (e.g., goals 1, 3, and 5 in the NDAA definition). Placing a computer in a role to act or choose is often qualitatively differently from using a computer to synthesize data. We refer to AI systems that, once operationalized, are designed to act or choose as *autonomous systems*. If, however, the objective is to synthesize information in order to support a human decision-making process, we refer to this as *algorithmically assisted decision making*.

Delineating between synthesizing information and acting on that information is key to the legal, moral, and ethical discussion of ML and AI. Fundamentally, using techniques to synthesize

data for personnel policy does not represent a dramatic break from traditional personnel management processes. ML represents a shift in the type or degree of information available. Information can be processed at a much more granular level, and syntheses can account for far more factors.

However, this more detailed information can continue to serve as just that — information to be weighed and assessed by a human responsible for making critical personnel decisions. The questions then are whether the algorithm's synthesis is fair and appropriate, and whether the human decision makers know how to adequately weigh the synthesis given its potential shortcomings. Appropriate training can help human decision makers in this process.[5] If computers are instead given the autonomy to make critical personnel management decisions, the types of legal, moral, and ethical questions can quickly compound — much as they have for other autonomous systems, such as autonomous vehicles.[6]

It may be far in the future before the automation of many personnel management decisions is even considered. However, some human deciders could effectively abdicate their decision-making role by uncritically accepting and acting on an algorithms' output. This underscores the need for effective training and safeguards for how ML outputs will be used, especially given the current trend of ML techniques increasingly entering the data-rich sphere of personnel management.

This chapter begins with a brief primer on ML, together with a short overview of some of its prominent strengths and weaknesses relative to more traditional statistical approaches. We then give a few vignettes for how ML may enhance personnel management. These vignettes provide specific contexts in which to consider the legal, moral, and ethical issues which will be addressed in later chapters.[7] We return to these vignettes in Chapter 6 for a more detailed examination of the issues at play.

A. Brief Primer on Machine Learning

ML algorithms are divided into a few broad classes. This includes **unsupervised** learning, **semi-supervised** learning, **supervised** learning, and **reinforcement** learning. A prominent

---

5 Depending on the nature of the decisions for which the information is being synthesized, a variety of methods (e.g., red teaming, tabletop exercises, analysis of alternatives) could be used in vetting and exploring how information might be used and determining the potential consequences. These methods might be used directly in training or they may also be used to assess how to effectively train decision makers.

6 As an example, suppose an autonomous vehicle suddenly detects a pedestrian. If the vehicle swerves, there will likely be an accident with another vehicle. If the vehicle does not swerve, the pedestrian will likely be hit. What should the vehicle do? Should it be deliberately programmed with a given reaction, perhaps by weighting the relative costs of damage? What testing should the vehicle undergo before it is sold to consumers? Who bears the liability of the outcome? The programmers, the manufacturer, the car owner, or another party?

7 For further examples of ways in which ML can enhance defense personnel management, see Julie Lockwood, Alan Gelder, Matthew Goldberg, et al., *Leveraging Machine Learning in Defense Analyses*, IDA Paper P-13174, (Alexandria, VA: Institute for Defense Analyses, May 2020).

difference between these classes of ML is the amount of information that is available about the outcome. Unsupervised learning algorithms do not have a specified outcome that the model is trying to predict. Instead these algorithms attempt to divide data into groups or clusters based on their underlying characteristics — a process which can be useful for tasks such as detecting anomalies.

Supervised learning, as mentioned earlier, tries to find the underlying patterns in data to predict a known outcome, such as whether a service member promotes to a certain pay grade or the class rank in which a cadet will graduate. Supervised learning relies on having an ample quantity of data where both the inputs and outcomes are known. However, in some cases, it can be difficult or costly to obtain or collect information on a given outcome.

Semi-supervised learning can be used when data on an outcome is available but limited. Reinforcement learning is a process for collecting and assessing data through experimentation. These methods are especially prominent in robotics, supply chain management, and strategy games (e.g., chess, checkers, Go), and other strategic situations where there are numerous possible choices.[8] Unlike supervised learning where the algorithm learns the link between a set of inputs and an outcome, reinforcement learning identifies the link between a choice, the feedback from that choice, and the environment in which the choice was made.

ML algorithms create models (or mathematical representations) of data. Like any model, an ML model should ideally be both **replicable** and **externally valid**. The model should be replicable in the sense that there should be a well-documented process leading to the creation of the model that can be replicated to produce the same model. This is critical for being able to audit and assess how the model was produced. The model should also be replicable in the sense that similar inputs produce similar outputs.

External validity refers to the extent that the model produces accurate outputs for data that were not used to create the model. For example, a model may be produced using data on individuals between the ages of 18 and 30 who are observed during the years 2010 to 2020. An initial test of external validity would be to assess the model's ability to produce accurate outputs for a similar group of individuals within the same age and time brackets.

At a minimum, to be externally valid, a model should produce reasonable outputs for comparable inputs. A more robust model would still produce reasonable outputs when the inputs change more dramatically — such as when using individuals from a different age bracket, or if the time period of observation changed substantially. Models that reflect human behaviors often need

---

[8] As an example, imagine a logistician who has to manage and select between many different shipping alternatives across numerous routes and transportation modes. The logistician has some information about shipping times, which can also be impacted by weather, the type of product being shipped, and other factors. Experimenting with alternatives can be costly since it introduces an element of uncertainty, but it also provides information that could improve the shipping network. Reinforcement learning provides an ML framework for balancing the cost of experimentation against the benefit of continuing to rely on known information.

to account for cultural, technological, social, institutional, and economic changes over space and time. Failing to do so can compromise external validity.

ML models may have poor performance if the training data is based on historical circumstances that are no longer representative. For example, since female service-members were not allowed to serve in combat positions historically, a model predicting retention (which may depend on the service-member's assigned job) may "learn" that females do not advance in combat positions, thus predicting a certain retention outcome that may be inaccurate with newer data.

If a model adheres too closely to its input data, it will underestimate natural variance that occurs in the real world and output **statistically biased** results when new data are evaluated. This is called **overfitting**. To reduce overfitting, supervised models are usually created using a subset of data, called the training set, and tested on a completely unseen set called the testing set. The success of a model should be identified through the accuracy of the outputted results from the testing set.

Machine learning algorithms prioritize **predictive power** over other statistical goals such as **interpretability** or statistical notions of efficiency, bias, or inference. For instance, modern machine learning methods tend to use highly complex models that lack the straight-forward interpretability of standard linear regression. A linear model expresses an outcome as the sum of the impacts of different inputs. The size of the impacts are free parameters, with each answering the question: how much would the outcome change, on average, if we saw an observation with more of this input while all other inputs remained the same?

Predictions using linear regression are straightforward: for each input, multiply the estimated impact times the quantity of that input, then add these products together. Slightly more complex linear models can account for non-linear relationships between an individual input and the output (e.g., the output tends to increase dramatically for smaller values of an input, but only gradually for larger values of the input). Another wrinkle of complexity is to capture interactions between inputs (e.g., the output decreases when two inputs have large values but increases when only one of the inputs has a large value). As the model incorporates more complex interactions and non-linear relationships, interpreting the impact that any single input has on the model becomes increasingly difficult. However, these complex interactions and non-linear relationships are central to many machine learning methods and are key to enabling more refined and nuanced predictions.

Another issue in any statistical model is **sampling variance**. In any given sample, part of the parameter estimate comes from the true impact of the variable and part of the estimate comes from idiosyncrasies in the sample.[9] Standard errors on parameter estimates come from the sampling

[9] That is, some things that impact the outcome are unobserved. In any given sample, observations with large values of an unobserved variable may also have large values of an observed variable (that is, the error term may be correlated with the covariates). In this case, linear regression will confuse the effects of the unobserved variable for effects of the observed variable. Correlations between the error term and the covariates will differ across samples, so parameters that provide a good prediction in one sample might not in another sample.

variance; they reflect how wrong parameter estimates might be given idiosyncrasies in the sample. Prediction using parameter estimates will partly come from true impacts (the "signal") and partly come from sampling variance (the "noise"). The part that comes from sampling variance makes the prediction worse (or "noisier").

If the true impacts are large and the sampling variance is small, then predictions from linear regression may be accurate. But if the linear model includes a large number of low-impact inputs, then the noise will overwhelm the signal, and the prediction will perform poorly. For this reason, researchers using linear regression for prediction will often restrict the model to including only those inputs that have a strong theoretical justification and are likely to have a high impact.

Regularization techniques play a central role in machine learning by adjusting parameter estimates to boost the signal-to-noise ratio. For instance, ridge regression (perhaps one of the earliest ML techniques) penalizes model complexity by biasing all parameter estimates towards zero. Although doing so reduces the signal, it reduces the noise to a greater extent; for a certain level of bias, a ridge regression prediction will outperform a linear regression prediction. Likewise, lasso regression drops inputs from the model (or, alternatively, sets parameters to zero) if the estimated parameters are too small. By doing so, lasso regression drops noisier inputs, which increases the signal-to-noise ratio in the prediction.

With these and other regularization techniques, researchers can add many parameters to the model without introducing too much noise. Consequently, researchers can include many more inputs, including those that might not have much signal. Furthermore, researchers can move away from linear models and towards more flexible non-linear models, such as tree-based models and neural nets, which have many more parameters than linear models. Such approaches allow researchers to include more data to catch more of the signal; regularization techniques prevent catching too much noise. Predictions using these techniques can be much more accurate than those using non-machine learning predictions.

The weaknesses of machine learning dovetail with its strengths. Most immediately, regularization techniques (and the nonlinear models that exploit them) reduce interpretability. Whereas parameters estimated in linear regression indicate the all-else-equal expected impact of additional inputs, regularized parameters are biased and not particularly meaningful in themselves. Similarly, the parameters in highly non-linear models do not speak to the impact of specific inputs on the outcome; rather, many parameters (and many different inputs) must be taken together to understand the impact of inputs on the outcome. This can be difficult to do.

Another weakness can derive from the ease of using a breadth of data in machine learning applications. When an analysis uses many variables, researchers and policy makers may be less aware of any risks from including certain variables or groups of variables. In non-machine learning analyses, researchers may be more selective about what data to include in the model, and consequently may be more likely to have considered potential drawbacks.

B. Examples of Personnel Management Processes

There are numerous potential uses of machine learning algorithms to improve personnel management processes. Organizations often maintain a variety of personnel records. This begins with resumes or other application materials with information on where people went to school, past employers, certifications, and other notable experiences and talents. Beyond resumes, there are records on job performance, assignments, individual and group efforts, training, evaluations, compensation records, retention information, promotion, and other measures of employee success.

These data can be used in a variety of applications, such as improving applicant screening processes by focusing on traits that are connected to successful performance; forecasting patterns associated with employee retention; understanding talent gaps; examining patterns in leadership development pipelines; and probing traits associated with job satisfaction. Many organizations are recognizing the value of personnel and human resource data in improving a spectrum of personnel management actions, and machine learning and other big data techniques are increasingly employed in this sphere.[10]

This chapter provides a few illustrative examples of potential uses of machine learning algorithms in personnel management processes. These fictional narratives illustrate capabilities that could be implemented now or in the near future. These technologies are not science fiction, but are now or will soon be implementable; thus, the pertinent questions are whether and how the technologies *should* be implemented. What guardrails are needed to avoid legal and ethical pitfalls, and what practices should be adopted to enable the responsible use and expansion of these technologies?

The specific machine learning use cases are hypothetical and presented for illustrative purposes.

---

[10] For instance, a 2013 *Forbes* article references many of the applications for people analytics in this paragraph. A 2016 *Deloitte* report noted that 77 percent of surveyed organizations globally believe people analytics is important, with smaller but growing numbers indicating that they were ready to incorporate people analytics within their organizations and develop predictive models. Examples include such things as predicting high-potential employees who may be a high-flight-risk, identifying working conditions that are associated with higher retention and productivity, analyzing external perspectives of the organization's culture through text analysis of various online postings, and using analytics on employee transactions to identify potential regulatory compliance breaches. See Josh Bersin, "Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age," *Forbes*, 17 February 2013, https://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/?sh=42cb4b2a4cd0; David Mallon, Jeff Moir, Robert Straub, "People Analytics: Gaining Speed," *Deloitte*, 2016, https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2016/people-analytics-in-hr-analytics-teams.html.

1. Fictional Vignette: Selecting Officers for Promotion

A handful of senior officers have gathered to participate in a selection board to promote a cohort of O-3 officers to the rank of O-4.[11] Each has taken the statutory oath to perform their duties on the board "without prejudice or partiality and having in view both the special fitness of officers and the efficiency of his armed force."[12] Acting without prejudice or partiality is a commitment that is not taken lightly. To help prevent undue influence, the members of the board have been given strict instructions not to discuss the board's deliberations with anyone outside the board. In fact, the Service Secretary is the only person who may appear in person to address the board on any non-administrative matter. All other information given to the review board must be in writing,[13] provided to each member of the board, and included in the board's record.[14]

As the selection board convenes, each senior officer receives a packet of information about each officer considered for promotion. This includes descriptions of duties, training and accomplishments, as well as supervisors' rating of the officer's performance, potential, character, and competencies.[15] This information has been compiled over the course of the promotion candidates' careers. Each candidate is responsible for ensuring the accuracy and completeness of his or her file prior to the board.

Given the number of candidates that the board must consider in limited time, board members find it impossible to digest and assess each candidate's entire file.[16] Pertinent pieces of information

---

[11] Five is the statutory minimum number of officers for a selection board (10 U.S.C. § 612 (a) 1). Members of the board must be an O-4 or above and in a higher rank than the rank for which officers are being considered.

[12] 10 U.S.C. § 613. This vignette is similar in style to a vignette presented in William Clayton, Joseph King, *Improving the Developmental Education Selection Board Process with Machine Learning*, IDA Paper NS P-13191, (Alexandria, VA: Institute for Defense Analyses, April 2020).

[13] In the Marine Corps, board members may not bring up outside information to criticize the candidate (such as from personnel interactions), but may bring up favorable information.

[14] DOD Instruction 1320.14, "DoD Commissioned Officer Promotion Program Procedures," 16 December 2020, p. 12–13. By law, members of the selection board cannot be coerced or reprimanded by their superiors in making their recommendations (10 U.S.C. § 616 (g)). The non-disclosure of board proceedings is further specified in 10 U.S.C. § 613a.

[15] The quantity of information in promotion packets can be quite large. Given time constraints, board members tend to rely heavily on certain fields within the promotion packets. In the Air Force, the senior rating officer may give a "Definitely Promote" (DP) rating, which strongly boosts board members' scoring. In turn, senior rating officers are subject to a quota of DPs to prevent them from handing these out too generously; senior rating officers that evaluate too few officers may not be permitted to give out any DPs at all. Board members sometimes speculate about whether the senior rating officer wanted to give a DP but was unable to due to the quota.

[16] In the Navy, a single board member reviews the promotion packet in detail, then briefs the other board members (who also have access to the promotion packets) on the major points. This procedure reduces the risk that relevant details will be overlooked, but increases the risk that a promotion outcome will be determined by the idiosyncrasies of the briefer.

on an officer are unintentionally overlooked. The criteria for evaluating officers may evolve throughout the process.[17,18] Human error[19] and fatigue take their toll.

Since 2013, DOD policy has explicitly permitted promotion boards to "consider automated computer summaries of information in an eligible officer's official military personnel record."[20]

To help provide consistency, each board member (in this fictional example) receives a handful of automated metrics summarizing different dimensions of the officer's performance and potential. Some of these metrics rely on machine learning. (Again, this is a fictional example; no military branches currently incorporate such machine learning based metrics into the promotions process.)

For instance, one metric assesses the probability that the O-3 being considered for promotion will successfully complete the key developmental assignments of an O-4.[21] Another metric assesses the captain's probability of qualifying for a command position years later. Other metrics assess the officer's probability of remaining in service for at least five more years, and whether the officer has critical skills for which the Service anticipates there will be a shortage in the next three years. This range of metrics provides the promotion board members with a common benchmark and supplements other heuristics, such as indicators from a senior rater, that promotion

---

[17] In the Air Force, each board member, without conferring with other board members, scores each promotion packet on a scale from one to ten. If the highest score exceeds the lowest score by more than two points (a "split"), then all board members discuss that promotion packet until the outliers move their scores such that no split remains. Splits become less frequent over the board's duration, reflecting that board members converge on similar criteria for evaluating packets.

[18] Navy promotion boards deliberately reconsider marginal promotion candidates over several rounds, known as the "crunch." Earlier rounds adjudicate the obvious candidates for promotion or non-promotion. Later rounds deliberately expose marginal candidates to a greater level of scrutiny.

[19] Promotion boards restrict available information in promotion packets so as to reduce the risk that board members will be overly or inappropriately influenced by certain types of information. For example, the Navy removed photographs from promotion packets in FY2006, reinstated photographs in FY2007, removed photographs in FY2017, and reinstated photographs in FY2019. Likewise, the Army removed educational credentials from promotion packets for lower ranked officers to prevent boards from relying too heavily on that information.

[20] See both the 11 December 2013 and 16 December 2020 versions of DOD Instruction 1320.14, Paragraph 3.2.c.(2)(a). The use of "automated computer summaries of information" is absent from the corresponding paragraph (6.1.3.2.1) of the preceding 24 September 1996 version of DOD Instruction 1320.14.

[21] DOD Instruction 1320.14 authorizes the use of "automated computer summaries of information," but the extent of what might be summarized may be controversial. For instance, is a summary limited to describing the attributes and performance of the individual? Can it also compare the individual with others who are being considered for promotion? Can it further compare the individual to previous cohorts, perhaps to promote common promotion standards over time? If so, can it consider the subsequent performance of previous cohorts, and compare the individual to those from previous cohorts who have met various performance standards? The metrics in this example take a broad interpretation of what may be summarized — well beyond interpretations used in practice today. Notably, without machine learning, some of these metrics would be difficult to compute, and they may not have been considered previously due to the infeasibility of summarizing such information. However, even if there are no technological barriers to computing these metrics, there is, of course, the underlying question of whether the metrics are appropriate to use at all.

boards have long used. The Service deliberately used multiple metrics to emphasize different dimensions of performance that may have been overlooked or difficult to quantify previously.

Designing these metrics raised several questions. What metrics should be produced? Are the outcome metrics reflective of traits that the Service values in its leaders, and can the outcome metrics be adequately measured? Also, how much should an individual's *projected* performance, based on the performance of others, factor into the decision-making process relative to the individual's *demonstrated* performance?

Too much reliance on projected performance may inadvertently close the door to talented individuals simply because they have an atypical background. This can effectively preempt an underdog from even having the chance to prove him- or herself on merit. On the other hand, to the extent that it is costly to gather and assess information on demonstrated performance, indicators of projected performance can help to provide a better signal of an individual's potential skills and quality.[22] Making decisions with inadequate information can be costly to the Service as an institution and to its members, but so can making decisions based on metrics that might exclude otherwise talented individuals. How might the Service appropriately balance these?

There are also other issues to consider, including: What data inputs are appropriate to use? Is the Service strictly bound to only use data found within an officer's promotion materials, or are other elements of the officer's official personnel record permissible? What elements in an officer's promotion materials are particularly salient for the current decision-making process, and how should salient and less-salient pieces of information factor into these metrics? What information, if any, does the Service need to give to officers about the metrics? How robust are the metrics across different populations, and how consistent are the metrics over time?

The Service also seeks to minimize the chance of bias in the promotion board's decision making. Consequently, by policy, board members may not be able to see the officer's race, ethnicity, or gender, and they may not be able to consider any of this information in their evaluation

---

[22] A subtle but significant issue here is whether the signal is something that the service members themselves have any control over. In contract theory, costly signals are used as a mechanism for allowing high-performing individuals to communicate their approximate skill level. For instance, some people may attend a prestigious university as a way to communicate to potential employers that they are particularly bright. Although others may be brighter or have received an even better education at a less reputable school, without the brand-name reputation of the prestigious university, it may be harder to quickly and credibly demonstrate their skill level. To the extent that individuals can internalize the costs and benefits associated with the choice of attending a prestigious university, using that type of signal may have different ethical considerations than using a signal that individuals have little or no control over. For instance, if service members hailing from colder climates like Minnesota are more likely to pass the cold-water portion of the Navy Seals training than service members from warmer climates like Hawaii or Florida, screening on the climate of the service member's hometown may be informative. However, since service members often have little control over the hometown they grow up in, screening on hometown may strike a chord of being unfair in a way that screening on an individual's alma mater may not.

of the officer's promotion potential.[23] Do these practices increase or decrease resulting diversity? What practices help to minimize bias in machine learning algorithms?

2. Fictional Vignette: Recruiting and selection for Special Forces

A recruiter for the Marine Forces Special Operations Command (MARSOC) looks on as a small group of Marines salute their commander on a humid day at Camp Lejeune, North Carolina. They have each just accepted an insignia badge featuring a gold eagle, wings stretched wide, with the words "Spiritus Invictus" on a banner above its head. They have just become Marine Raiders. Selection into this elite role is an achievement that can be years in the making.

Today's ceremony celebrates the unconquerable spirit of the Raiders before him, but the recruiter also feels a sense of personal pride. Since the Marine Corps established MARSOC in 2006, this year marks the highest-ever graduation rate from the program's notoriously exacting selection and training process. The high graduation rate does not represent a compromise of standards. Rather, these recruits performed at a much higher level. The recruiter (in this fictional example) attributes this to a novel use of analytics in the recruitment of potential Special Forces candidates.[24]

The Special Forces require candidates to complete an assessment and training process that includes an exacting battery of physical, psychological, social, and mental aptitude tests. The high standard typically corresponds to a high failure rate. This ensures that the Marine Corps selects the best candidates to join these elite units, but the process also entails considerable expense in terms of time, staff, materials, and the propensity of candidates who failed the testing to remain in the military. By better identifying, recruiting, and quickly screening promising candidates up front, the Marine Corps can reduce these costs and improve its return on investment.

In years past, MARSOC recruiters (in this fictional example) targeted information about the program to Marines who met core prerequisites: high scores on the general technical section of the Armed Services Vocational Aptitude Battery, fitness tests, and swim assessments; the ability to meet medical criteria; and a clean disciplinary record.[25] Conventional wisdom held that certain

---

[23] See, for example, Department of the Army, MILPER Message Number 20-209, "Elimination of Department of Army (DA) Photos, and Race, Ethnicity and Gender Identification Data for Officer, Warrant Officer, and Enlisted Department of the Army Centralized Selection Boards," 8 July 2020, https://usasd.armylive.dodlive.mil/files/2020/07/MILPER-MESSAGE-20-209-REMOVAL-OF-DA-PHOTO-FROM-SRB.pdf

[24] Again, the specific machine learning use cases here are hypothetical for illustrative purposes.

[25] "Becoming a Critical Skills Operator," Marine Forces Special Operations Command Website, last accessed May 2, 2022, https://marsoc.com/career-paths/critical-skills-operator/.

specialties, such as reconnaissance, would best prepare a Marine to succeed in the Special Forces, so recruiters focused their efforts there.[26]

Recently (in this fictional example), recruiters shifted their efforts based on information they gained from a machine learning model that identified attributes most predictive of success in the Special Forces program. The model uses extensive data on prior Special Forces candidates and members to analyze not only predictors of graduation from the training program, but also predictors of long, successful Special Forces careers.[27]

The algorithm considers extensive personnel records, including details from previous Fitness Report evaluations, accession records, academic records, and other test, evaluation, and demographic data on file. This algorithmic approach enables recruiters to scale their efforts in a sense by considering high-potential Marines from across the entire Corps instead of the limited set of communities that they typically spend most of their time with.[28] Recruiters targeted their efforts toward Marines that the model predicted had at least a 75 percent probability of successfully completing all portions of the selection and training process. However, recognizing some limitations in the model's forecasting capabilities, recruiters still considered other candidates with promising attributes.

Consequently, the cohort of Marines they sent to this year's Assessment and Selection course (in this fictional example) had a higher chance of success, statistically speaking, than any other.[29]

---

[26] Most common specialties based on https://www.marinecorpstimes.com/news/your-marine-corps/2019/04/11/marine-officers-are-more-successful-during-raider-selection-but-marsoc-is-fielding-raiders-at-a-steady-rate/

[27] As with the previous fictional example, there is an underlying question of how much an individual should be evaluated on objective measures of the individual's *demonstrated* performance as opposed to measures of the individual's *projected* performance, as forecasted by the performance of others with similar characteristics. Too much focus on projected performance may overlook key talents. It could also create feelings of disaffection in the force with people viewing the system as unfair or based overly on factors outside the individual's control.

[28] There are some tradeoffs here. For instance, the algorithmic prediction may enable the recruiters to gain some information about a broader population along the dimensions for which this metric is attuned. However, there will likely be Marines that this metric will not adequately measure who could easily be overlooked by recruiters who place too much emphasis on the metric. An analogy here is credit scores: those who borrow money and pay it back responsibly will tend to have high scores, those who borrow and have trouble paying it back will tend to have low scores, but those who do not borrow will largely be missed by these scores – even if they routinely pay their rent and other bills on time and practice other financially prudent patterns, such as saving for a rainy day, saving for retirement, and being properly insured. These other measures of financial responsibility are not included in traditional credit scores (unless, perhaps, as a negative signal of a grossly delinquent customer that a landlord or utility company bothered to report). Consequently, there is a portion of the population who could credibly demonstrate that they are financially responsible through other metrics but who are still excluded from loan considerations by major lenders.

[29] It is possible that focusing too much on the objective of identifying a cohort that has a high probability of completing Assessment and Selection may detract from other meaningful objectives. Algorithmic predictions of success might, for example, inadvertently produce a cohort that does not have a very diverse set of backgrounds

Still, their success was not guaranteed. Each Marine still had to meet the same rigorous standards in Assessment and Selection in order to advance to the nine-month Individual Training Course, which is the final hurdle to becoming a Raider. Previously, only 20 percent of the enlisted Marines who started this multi-phase selection and training process ultimately became a Raider.[30] This year, the rate increased to 45 percent.

Looking to next year (in this fictional example), MARSOC leaders are considering two plans to further improve the efficiency of this process. One plan is to use a similar machine learning model to identify core elements of the cognitive, psychological, physical, and leadership tests administered during the training that are particularly predictive of ultimate success. Using these predictors, MARSOC plans to develop a one-week abbreviated selection program that they can use to screen prospective recruits up front.

A short but effective screening process would reduce the cost of failure to both the Marine Corps and to the individual Marine. Depending on how effective the abbreviated selection program is at screening, individuals may advance to a shortened Assessment and Selection course or even directly to the Individual Training Course. The other plan is to hone the recruiting model to identify individuals who can advance directly to the Individual Training Course. Being offered a direct pass to the Individual Training Course could provide the top echelon of high-performing Marines (who are sought for by other elite programs as well) with an additional incentive to join MARSOC. This option may be limited, for instance, to Marines that the model predicts to have greater than a 95 percent probability of successfully completing all portions of the selection and training process.[31]

As MARSOC irons out the details for these plans, it has some concerns. Primarily, it does not want to compromise on the exacting standards of those who qualify as Raiders. Do those who are predicted to have a high probability of successful completion truly have the needed skills and attributes? How much is lost by not having promising recruits complete the full Assessment and Selection process, such as the lasting impacts of bonding and team-building that would be difficult

---

and experiences. Such a lack of diversity may result in a fighting force that has a more limited set of skills on the battlefield. It may also lead to a narrow-minded culture that in turn becomes detrimental to recruiting efforts.

30 Between FY2014 and FY2018, only 28 percent of the enlisted Marines who began MARSOC's Assessment and Selection successfully completed it; the passing rate for officers is higher at 46 percent. Those that are successful then advance to the Individual Training Course, where the average completion rate for both enlisted and officers between FY2013 and FY2018 was 73 percent. Thus, the start-to-finish completion rate is 20 percent for enlisted Marines and 34 percent for officers. (Shawn Snow, "Officers are more successful during Raider selection, but MARSOC is fielding Marines at a steady rate," *Marine Times*, 11 April 2019, https://www.marinecorpstimes.com/news/your-marine-corps/2019/04/11/marine-officers-are-more-successful-during-raider-selection-but-marsoc-is-fielding-raiders-at-a-steady-rate/)

31 A caveat here is that predictions would be made using data from prior cohorts. If the skills, attributes, or characteristics that MARSOC values change over time, and if its training changes accordingly, then predictions of success would be based on prior training requirements and not necessarily on what MARSOC values today.

to produce otherwise? MARSOC is also concerned about diversity, equity, and inclusion. How can they ensure that the recruiting algorithm and the abbreviated selection program are adequately inclusive? Finally, MARSOC wonders how much information about its algorithms it is legally or ethically obligated to share with recruits or the broader Marine Corps community.

3. Fictional Vignette: Better programming for training slots

Air Force leaders in the Reserve and Guard have been frustrated with long-vacant slots in critical career fields. In some cases, the cause is not a lack of recruits, but long wait times to get individuals through the required training pipeline.[32] It is also not just a matter of scheduling. Accurately budgeting for training slots requires anticipating training needs a year or more in advance of the actual need. Reserve and Guard elements (in this fictional example) have attempted to forecast demand for training seats based on data that a team of Career Field Managers receives from various sources. This data, however, is rarely timely and often flawed, resulting in unreliable estimates.

One promising solution is to forecast when individuals who are currently serving are likely to leave. By better anticipating outflows, and by knowing the staffing requirement needed in each skill set, the Reserve and Guard can identify the needed number of incoming individuals who require training. Accurately forecasting the number of individuals expected to leave each quarter over the next few years would enable better budgeting and planning for training needs. It may also help inform recruiting efforts.

A branch of statistics called survival analysis enables researchers to study when a process — such as serving in a given career field in the Reserve or Guard — is likely to end. Classical tools answer this question to some degree, but machine learning adds greater flexibility and uses large data resources more effectively. In this case, extensive personnel data spanning multiple decades can be used to uncover detailed patterns for when individuals with certain career histories and other characteristics are likely to exit service.

This adds a rich level of fidelity to the forecasts across various Reserve and Guard locations and positions, numerous Air Force Specialty Codes, and other skill profiles. Once implemented in a dashboard, Career Field Managers and others could break down detailed forecasts to determine likely training needs. The dashboard masks details at the individual level and instead focuses on aggregated forecasts by Air Force Specialty Code, geographic region, and other needed fields.

In announcing the rollout of this new process, leaders faced some push back (in this fictional example). Some recruitment and training managers were not convinced that a machine learning algorithm could do a better job of anticipating needs — especially one that focused on attrition

---

[32] See, for instance, Scott Maucione, "As cyber units expand, National Guard has training backlog," *Federal News Network*, 16 March 2016, https://federalnewsnetwork.com/defense/2016/03/cyber-units-expand-national-guard-training-backlog/.

instead of accession. Who would be accountable if the algorithm did a worse job of predicting training needs? Would current methodologies be maintained as a fallback, at least until the model's performance could be consistently demonstrated?

The model would also need to be validated to ensure it was producing meaningful forecasts across multiple forecast horizons and populations. Leaders (in this fictional example) do not foresee any potential legal, moral, or ethical issues for using retention and attrition forecasts; although individual-level data are used, forecasts are aggregated by Air Force Specialty Code and other fields that are needed for budgeting and planning purposes. Leaders do, however, want to ensure that the model is properly vetted.

4. Fictional Vignette: Targeted retention interventions

An Army Captain sits deep in thought, considering her next move. At this point in her career, she planned to leave the Army, and she recently began working on applications for Master of Business Administration (MBA) programs. Between her 3.9 GPA at West Point and eight years on Active Duty, including two deployments to the Middle East and a successful performance as a company commander, she feels good about her chances of acceptance into a top-ten program.

Today, however, she received an email from Human Resources Command that gave her pause. The email offered her a sizeable bonus and preferential eligibility to apply for a fully-funded 18-month master's program at a military college in exchange for committing to five additional years of service. She is intrigued by the idea — it would be nice to avoid six figures of student debt — but she is also unsure about the prospect of making such a commitment.

If she served for another five years, she would be just seven years away from the additional retirement benefits that accrue after 20 years of service. After five years, she may be inclined to continue her service all the way to retirement. This feels like a tipping point between a full military career and the plans she had been cultivating.

The timing of the email also seems uncanny. No one had mentioned receiving a similar offer to her, and she thought she would have heard some news if the Army was conducting a sweeping retention initiative. She had told no one but her parents about her plans to leave the Army or her interest in a master's degree. "How did they know?," she wondered.

In fact, the Captain (in this fictional example) received the email as a result of an Army initiative that used a machine learning algorithm to identify high-potential officers who are likely to leave military service in the next two years, with the goal of retaining them through customized incentives. The algorithm was not eavesdropping on private conversations. Rather, it was analyzing an extensive body of Army personnel data, including career history, performance information, family status, skill inventories, and other administrative records.

It also considered external data, such as job market trends. Combining this information with exit patterns of similar individuals who have left military service, the algorithm output person-level forecasts for an individual's propensity to leave military service. These person-level forecasts

were then used to target retention interventions such as the one the Captain is now considering. The Captain is, of course, free to reject the offer with no repercussions.

The Army (in this fictional example) does not automatically extend retention incentives to those that the machine learning algorithm predicts have a high probability of leaving in the next two years. That prediction is combined with information about forecast shortages and surpluses for given skills, as well as with information about the officer's record of performance. Individuals at Human Resources Command review this information and decide when to extend retention incentives, using a standardized rubric to guide their decisions. However, the Army is concerned that high performers who are not predicted to exit in the near future may feel underappreciated. The Army has therefore decided (in this fictional example) not to broadly publicize the program. They will instead keep the program relatively small and only offer a limited number of carefully targeted incentives.

Some leaders question the opaqueness of this approach, thinking it will exacerbate retention problems when soldiers find out that colleagues with comparable service records were offered incentives not available to them. Others counter that other high-performing officers may change their behavior to try to appear as if they were likely to exit if they knew that was a determinate for being considered. Others want to frame the retention program as merely a trial program until the Army can assess the effectiveness of these incentives over multiple years. There are also questions about the underlying algorithm. Are there potential biases in retention patterns in the past that may introduce bias into the set of officers that are being flagged as likely to exit, and therefore as a possible recipient of a retention incentive? Are there data that should not be included in the algorithm?

This page is intentionally blank.

3. Moral and Ethical Framework

The proliferation of machine learning, artificial intelligence, and "big data" have opened a broad dialogue on moral and ethical uses of algorithms and data. As new and unforeseen moral and ethical questions arise, it can be helpful to assess these questions from foundational principles. We set the stage with an overview of normative ethics — the branch of philosophy which describes and studies theories of moral behavior. We then explore ethical lessons and applications within the machine learning context.

A. Moral Philosophy: Normative Ethics

Morals and ethics both pertain to philosophies or standards of right and wrong. Morals may be considered more personal than ethics and subject to more individual interpretation. A person's morals may become evident through their convictions of what they may or may not do, regardless of how societally acceptable the action may be. Ethics, by contrast, may be understood to reflect collectively held morals that are often invoked as universal standards in a given context. Ethics are typically applied to a given context as a code of behavior within a profession (e.g., business or medical ethics), religion (e.g., Protestant or Catholic ethics), or other groups.[33]

Normative ethical theories seek to define standards for moral behavior by classifying certain phenomena as "right" or "wrong."[34] These theories typically prescribe a set of principles for evaluating the morality of a behavior or other phenomenon. Theories of normative ethics are generally grouped into three major approaches, focusing either on the consequences of an action (consequentialist ethics); the correctness of an action, independent of what the consequences may be (deontological ethics); or the goodness of the person doing an action (virtue ethics). provides a brief summary of these three approaches, along with some of the major theories that fall within each.

---

[33] "Ethics vs Morals: Is there a difference?," Merriam-Webster Website, last accessed May 2, 2022, https://www.merriam-webster.com/dictionary/ethic#note-1. See also "What's the Difference between Morality and Ethics?," Britannica Website, last accessed May 2, 2022, https://www.britannica.com/story/whats-the-difference-between-morality-and-ethics.

[34] "Ethics and Contrastivism," The Internet Encyclopedia of Philosophy Website, ISSN 2161-0002, last accessed May 2, 2022, https://iep.utm.edu/ethics/#H2.

Table 1. Normative Ethical Approaches and their Major Theories

Consequentialist Ethics	An action's consequences determine its morality
Utilitarianism	Actions that maximize good consequences are moral
Ethical egoism	Actions that maximize the agent's self-interest are moral
Ethical altruism	Actions that maximize the benefit to others are moral
Deontological Ethics	Compliance with rules determines the morality of the action
Kantian theory	People have a categorical imperative to follow moral principles
Duty-based theory	People have obligations to act or not act in certain ways
Rights-based theory	People have inherent rights which others must not violate
Contractarianism	People accept moral principles as part of a social contract
Virtue Ethics	The fundamental character traits of virtue, moral wisdom, and fulfilment constitute a person's morality
Eudaimonic	Living with virtue leads to "eudaimonia," or the greatest good
Agent-based	An action is moral if the intentions of the agent are virtuous
Target-centered	An action is moral if it fulfills the purpose that a virtue aims to achieve

Consequentialist theories assess the morality of actions based solely upon the resulting outcomes or effects of those actions.[35] Actions are moral if they result in positive consequences or contribute to some defined "good" (although consequentialist theorists disagree on what constitutes "good"). Recognizing that an action may lead to some good and bad outcomes, classic utilitarian philosophers claim that an action is moral if the net effects of the action are good. Morality therefore becomes a calculation that weighs the magnitude and severity of good and bad consequences. Variations of utilitarianism differ in their level of analysis; "act" utilitarianism evaluates the consequences of each individual action, whereas "rule" utilitarianism evaluates the consequences of rules.[36]

---

[35] See, for example, Walter Sinnott-Armstrong, "Consequentialism", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

[36] "Ethics and Contrastivism," The Internet Encyclopedia of Philosophy Website, ISSN 2161-0002, last accessed May 2, 2022, https://iep.utm.edu/ethics/#SH2c.

Other consequentialist theories differ from utilitarianism in how they calculate the net good. For example, utilitarianism balances all consequences, but ethical egoism considers morality only in terms of the consequences for the individual performing the action, and ethical altruism considers morality only in terms of consequences to others.

Deontological ethics are the converse of consequentialist ethics. Rather than assessing morality in terms of the consequences of an action, deontological ethics assess morality based on the action itself.[37] Actions that comply with moral norms are moral, regardless of whether the effect of the action is good or bad. The nature (and sources) of moral norms varies among deontologist theories. Duty-based theories hold that individuals have both permissions and obligations to act or not act in certain ways, while rights-based theories hold that individuals have rights which others must not infringe upon.

The theories of philosopher Immanuel Kant exemplify the core characteristics of deontological ethics. Kant presented the idea of a "categorical imperative," or an unconditional moral obligation to act according to formal principles. In Kant's view, such principles derive from a test of whether a particular rule of conduct could serve as a universal law that applies equally to everyone. If everyone could conceivably comply with a universal law, and if compliance with the universal law would be rational, then that conduct is moral.[38] Kant's views require that morality be determined *a priori* (i.e., through theoretical reasoning rather than empirical observation) and apply in all contexts.

Other schools of thought within deontological ethics build upon Kant's thinking, particularly in terms of how they arrive, *a priori*, at moral principles. Contractarian theories hold that moral conduct should be assessed according to principles that individuals in a society accept as being part of a social contract. Thomas Hobbes, a major philosopher of contractarianism, argued that individuals will accept moral principles that maximize their self-interest, and that result from balancing cooperation with, and protection from, the interests of other members of society.[39]

The third area of normative ethics, virtue ethics, conceptualizes morality in terms of a person's fundamental character traits. Virtue ethicists look to the concepts of virtue, wisdom, and "eudaimonia" — a sense of enduring and fulfilling happiness or joy.[40] Virtue here extends beyond considering one's duty (in a deontological sense) or the consequences of one's actions (in a

---

37 See, for example, Larry Alexander and Michael Moore, "Deontological Ethics", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2020/entries/ethics-deontological/.

38 Robert Johnson and Adam Cureton, "Kant's Moral Philosophy," *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2019/entries/kant-moral/.

39 Ann Cudd and Seena Eftekhari, "Contractarianism," *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2018/entries/contractarianism/.

40 See, for example, Rosalind Hursthouse and Glen Pettigrove, "Virtue Ethics," *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/.

consequentialist sense). Rather, it is a complex state of being in which a person is motivated by intrinsic values and carries out those values in all aspects of life.

Moral wisdom refers to the practical application of virtue, or the ability to recognize the contexts in which upholding one's values is beneficial or harmful. For example, honesty is a virtue, but moral wisdom is knowing when and to what degree to practice honesty. Eudaimonia is an Aristotelian concept of the greatest good — a good that is sought for its own sake. It embodies the idea of a life well lived and the resulting fulfillment that brings, although virtue ethicists tend to disagree on its role and importance.

Eudaimonist theories of virtue ethics consider eudaimonia to be justification for living virtuously. In this view, an individual who develops and applies virtue and moral wisdom has a eudaimon life. This conceptualization separates these character traits from external factors, including the consequences of an individual's behaviors. Other forms of virtue ethics do not give eudaimonia such a central role. Agent-based theories consider an action to be moral if the motivations and intentions of the agent are virtuous, or if the action is consistent with what a person with virtuous motivations would have done. Virtues may thus be identified by observing the behaviors and dispositions of admirable (or, by contrast, deplorable) individuals in society.

Target-centered forms of virtue ethics consider an action to be moral if it realizes the same aim or "target" that a virtue is meant to achieve. For example, if the virtue of generosity fundamentally targets improving the well-being of others, then an act that "hits this target" is virtuous. Thus, whereas eudaimonic theories place primacy on innate character in assessing the morality of an action, agent-based and target-centered theories leave more room for the possibility of virtuous behavior by non-virtuous individuals. Overall, virtue ethics emphasize the process of building virtue, moral wisdom, and eudaimonia.

B. Normative Ethics and Military Personnel Policy

The consequentialist, deontological, and virtue ethics framework each provide a meaningful approach for assessing military personnel policy. Each has useful insights regarding the moral and ethical implications of using machine learning in military personnel management. Yet, each also has weaknesses. In assessing the appropriateness of a given issue, each framework may arrive at the same conclusion, but they may justify the conclusion in different ways. Or they may disagree entirely.

Given that competing views and priorities may arise during such deliberations, it is our view that these deliberations are essential to promoting ethical decision-making regarding the use of machine learning in personnel policy.[41] To inform these deliberations, decision-makers may

[41] Theorists of deliberative democracy argue that by facilitating collective decision-making based on consideration of competing views and interests, such deliberation is the most justifiable approach to resolving conflicting moral views. E.g., Amy Gutmann and Dennis F. Thompson. "What Deliberative Democracy Means," in *Why deliberative democracy?,* Princeton University Press, 2009.

consider how DOD priorities and the values embodied in U.S. institutions intersect with the three normative ethical approaches.

A driving priority for DOD personnel policy is maintaining the military readiness and lethality of an all-volunteer force (the ethics of military and lethal force is an entire topic in itself). In the context of consequentialism, this priority could be considered the "good" which DOD actions must aim to maximize. Recognizing that there are a multitude of actions that contribute to this good, and recognizing that certain actions may have conflicting effects or undesirable side-effects, this approach furthermore requires an assessment and valuation of the various benefits and harms that may result from a policy or practice.

The course of action that most contributes to military readiness, while also upholding commonly espoused American values and doing the least harm to other outcomes, would be considered moral. The challenge is in ascribing appropriate weights to each benefit and harm across each dimension that the DOD or broader U.S. public cares about.

For example, until recently, the military services had policies that prevented Sikh service members from wearing the beards and turbans that are prescribed by their religious faith.[42] A consequentialist view focused solely on military readiness may weigh enforcing uniform standards for dress and grooming as a positive factor that contributes to readiness. However, this must be weighed against the cost to readiness from excluding Sikhs from the recruiting pool who would have served (or served longer) in the all-volunteer force but choose not to due to the military's prohibition on wearing of beards and turbans for religious purposes. If the positive factor outweighs the negative factor, then the policy would be considered moral.

Yet, this limited calculus ignores U.S. laws requiring reasonable accommodations for religious practices.[43] Thus, the end goal of maximizing readiness must be balanced against other values. A consequentialist view that is overly focused on a single objective (or which does not adequately balance multiple objectives) may not be a sufficient measure of morality. This suggests that deontology, too, has useful applications.

Deontology provides several lenses through which to view the morality of military personnel management practices, each of which centers around rules and norms defined *a priori*. The deontological concept of "duty" has clear manifestations in a DOD context: Service members and

---

[42] Following *Singh v. McConville* (187 F. Supp. 3d 152, D.D.C. 2016), the Army updated its policy on dress and grooming to accommodate the wearing of turbans, beards, and hijabs for religious purposes (Army Regulation 670-1, 25 May 2017). The Air Force and the Department of the Navy followed suit by updating their policies in 2020 (see Air Force Instruction 36-2903, 7 February 2020, and Navy BUPERSINST 1730.11A, 16 March 2020).

[43] To balance the sometimes-competing objectives of laws and religious practices, the Religious Freedom Restoration Act of 1993 (Pub. L. No. 103-141, 107 Stat. 1488) requires federal laws to make religious accommodations unless the law advances a "compelling government interest" and is also "the least restrictive means" for furthering that interest.

DOD leaders take an oath to support and defend the U.S. Constitution.[44] Promoting military readiness could therefore be considered to directly contribute to this duty to defend the Constitution and the country it represents. In a duty-based approach, personnel management practices that comply with these duties would be considered moral.

Rights-based theories would caution that these practices must not infringe upon the rights of individual service members. The preservation of individual rights is consistent with Jeffersonian values of equality before the law and equality of opportunity, as enshrined in the Declaration of Independence, the Constitution, and the Bill of Rights.[45] Perhaps more salient to military personnel management is the contractarian idea of the social contract that these founding American documents also embody.[46] Given that the U.S. military consists of an all-volunteer force, joining the military can be seen as entering into a social contract wherein the new recruit expects that his or her interests will be balanced with those of the military and the country that he or she serves. In this view, personnel practices that uphold this contract are moral. For example, supporting the career development of service members could be considered moral, and so could the out-processing of members who detract from readiness.

Virtue ethics holds that individual character is essential to morality. This approach would argue that the military leaders and those involved in personnel management decisions should possess virtuous characteristics and intentions. They should take actions that a virtuous decision-maker would make. For instance, the DOD attempts to encourage these ideas during the promotions process: it counts on members of promotion review boards to carry out their review in a fair and unbiased manner. "Fairness" is a virtue in this context.

The military also seeks to build character and moral wisdom in service members themselves — an idea that is embodied in the Warrior Ethos, Service creeds, and principles of leadership. To the extent that the culture and values of the military inspire service members to develop and live by virtuous characteristics, the actions of those service members can be viewed as moral. From an agent-based approach, the actions may be moral because the service members' intentions are

---

[44] The oath for those "elected or appointed to an office of honor or profit in the civil service or uniformed services" is in 5 U.S.C. § 3331. The oath for those "enlisting in an armed force" is in 10 U.S.C. § 502. Both oaths contain the phrase: "I will support and defend the Constitution of the United States against all enemies, foreign and domestic."

[45] For instance, the Declaration of Independence refers to an entitlement of self-government derived from "the Laws of Nature and of Nature's God," as well as the "self-evident" truths "that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." The U.S. Constitution was designed to "secure the Blessings of Liberty to ourselves and our Posterity." The egalitarian prerequisites for holding public office were limited to age, citizenship status, and residence within a state. Importantly, "no religious Test shall ever be required as a Qualification to any Office or public Trust under the United States." The Bill of Rights likewise enumerates several rights which are to be held inviolate.

[46] The Declaration of Independence refers to government as a social contract: "Governments are instituted among Men, deriving their just powers from the consent of the governed." The Constitution likewise captures the idea of a social contract with its opening pronouncement of "We the People of the United States."

themselves virtuous. Or from a target-centered approach, an action may be moral because it was the action that a virtuous person would take.

C. Applied Ethics: Ethical Principles from ML and AI applications

Whereas normative ethics proposes broad definitions of moral behavior, applied ethics considers the practical application of morals in the context of a specific field or issue. Applied ethics are critical for bridging the gap between the abstractions of moral philosophy and actual applications. In an extensive volume on ethics in autonomous and intelligent systems, the Institute of Electrical and Electronics Engineers (IEEE) advocates for "the critical-thinking terminology of philosophers, policymakers, and other stakeholders" on autonomous and intelligent systems to "be translated into norms accessible to technicians."

Philosophical ideals need to be accessible. Illustrative examples can link normative ethics concepts with concrete applications for machine learning and artificial intelligence.[47] Others have also developed frameworks for translating philosophical and societal values into practical principles for ethical machine learning and artificial intelligence. This section briefly explores applied ethics for ML and AI.

An initial step in any applied ethics framework is to identify the ethical dimensions of the issue at hand. The IEEE notes that the process of embedding human values into the design of algorithms used in human decision-making is particularly prone to ethical considerations. Since algorithms do not have their own morality, the design of algorithm—and assumptions about the dependability of the data which feed them — represent both explicit and implicit ethical presumptions. These presumptions are then embedded in the decisions or actions that are based on these algorithmic outputs. Mittelstadt et al. (2016) propose six sources of ethical challenges in ML and AI.[48] These sources include:

- *Inconclusive evidence*: algorithms form predictions of probable, not certain, outcomes, and these predictions are based on correlations, not causation, in the data. As such, there is a risk of algorithms making incorrect predictions.
- *Inscrutable evidence*: the source, scope, and quality of the data an algorithm processes, and how the algorithm translates the data into a prediction, may be unknown. This may pose challenges for the scrutiny of the algorithm and its predictions.
- *Misguided evidence*: algorithms may use data that contain quality issues or bias. This may result in unreliable or biased predictions.

---

[47] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019, pg. 47. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

[48] Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3, no. 2 (2016).

- *Unfair outcomes*: the predictions an algorithm produces may lead to decisions or actions that have undesirable discriminatory effects, even if the algorithm used sound evidence to produce the prediction.
- *Transformative effects*: by generating new and potentially unexpected insights, the use of algorithms can cause changes in social and political organization, as well as in the ways in which humans view the world.
- *Traceability*: it may be difficult to identify harms algorithms cause, the sources of those harms, and the individuals to hold responsible for them.

The above list is certainly not exhaustive. For instance, machine learning applications will tend to favor the use of inputs and outcomes that can easily be quantified. However, this tendency can inadvertently (and perhaps detrimentally) shift the focus of analysis. The machine learning application may be answering the wrong question. For instance, there are natural dangers to personnel systems that rely too much on easily quantifiable outcomes. A military officer may have an impressive service record but a toxic leadership style that is hard to quantify. Machine learning algorithms — like some manual selection processes — can fixate on measurable aspects of a service record but neglect less quantifiable personality characteristics and behaviors that play important roles in leadership positions.[49]

AI stakeholders have taken consequentialist, deontological, and virtue ethics approaches to addressing these ethical questions.

1. Consequentialist approach to AI ethics

Consequentialist perspectives have been used to justify the use of ML and AI broadly. Broadly speaking, this view holds that as long as the benefits of AI outweigh any harms or risks, then the use of AI is ethical. A corresponding argument is that *not* using AI may potentially be unethical if its benefits outweigh the drawbacks of not using it. The scope of what should be counted as benefits and risks can vary across different schools of thought.

Utilitarianism takes an expansive view, encompassing risks and benefits that may impact social justice, short- and long-term societal and personal effects, economic consequences, and other effects for humanity broadly.[50] The application of this approach may be achieved by training models to discern between good and bad consequences. A challenge here is properly defining, measuring, and balancing all the relevant dimensions that may impact an assessment of good and

---

[49] To overcome the tendency to assign command positions to those who look good on paper, the Army recently implemented its Command Assessment Program to more holistically assess officers' suitability for command. See, for instance, Everett Spain, "Reinventing the Leader Selection Process: The U.S. Army's new approach to managing talent," *Harvard Business Review*, November-December 2020.

[50] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019, pg. 39. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

bad consequences. To the extent that can be accomplished, ethical AI can be achieved by programming algorithms that maximize some form of social objective (assuming an appropriate objective can be defined and quantified).[51] The utilitarian framework also suggests that models may use rules that maximize good outcomes on average rather than calculating the effects of every decision.

2. Deontological approach to AI ethics

Others have perceived a need for a more deontological approach to AI ethics. To guide and govern the developers and users of ML and AI algorithms, they have set out ethical principles and frameworks that define *a priori* universal principles. With the expanded use of ML and AI technology in recent years, there is an active dialogue to define and refine these principles. For instance, Stanford University's *AI Index 2021 Report* noted that there were over 3,000 news articles on AI ethics, data privacy, and algorithmic fairness and transparency that were published globally in 2020.[52] The outputs of this discourse can be seen in the release of AI ethics frameworks by various consortiums, industry leaders, think tanks, and governments.

One systematic global review of AI ethics guidelines identified 84 major documents that lay out soft law and non-legal guidelines for ethical AI.[53] This list includes such documents as the American Medical Association's "Policy Recommendations on Augmented Intelligence in Health Care H-480.940;" IBM's "Everyday Ethics for Artificial Intelligence: A Practical Guide for Designers and Developers;" and the European Union High-Level Expert Group on Artificial Intelligence's "Ethics Guidelines for Trustworthy AI." Of the 84 documents, U.S. entities released the largest number (21), followed by European Union entities with 19. The majority of all of these documents were released after 2016.

Despite the diversity of organizations and nations that have released guidance for the development and application of AI, the frameworks frequently overlap in the principles they

---

[51] Bill Hibbard, "Ethical artificial intelligence." arXiv preprint arXiv:1411.1373 (2014).; William Bauer, "Virtuous vs. utilitarian artificial moral agents." *AI & SOCIETY* 35, no. 1 (2020): 263-271.

[52] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, et al., "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021, p. 131. https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf. The report also noted a tripling of AI references in the 116th U.S. Congress (2019–2020) over the 115th U.S. Congress (2017–2018), which itself had a ten-fold increase in AI references over prior years (p. 171).

[53] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1, no. 9 (2019): 389-399. An even more expansive collection of AI ethics guidelines is the *AI Ethics Guidelines Global Inventory*, which currently catalogs more than 160 distinct guideline documents (see https://inventory.algorithmwatch.org/).

propose. Table 2 displays this overlap, drawing from three recent reviews of AI guidelines and ethical frameworks.[54]

Table 2. Top 10 ethical principles in AI guideline documents, by review

	AI Index Report 2019 59 documents reviewed	Hagendorff (2020) 22 documents reviewed	Jobin et al. (2019) 84 documents reviewed
1	Fairness	Privacy protection	Transparency, explainability
2	Interpretability, explainability	Justice, fairness, non-discrimination	Justice, fairness
3	Transparency	Accountability	Safety, security, non-maleficence
4	Accountability	Transparency, openness	Accountability, responsibility
5	Data privacy	Safety, cybersecurity	Privacy
6	Security, reliability, robustness	Common good, well-being, sustainability	Beneficence, well-being, social good
7	Human control	Human oversight, control, auditing	Freedom, autonomy, consent
8	Safety	Solidarity, inclusion, social cohesion	Trust
9	Diversity and inclusion	Interpretability, explainability	Sustainability
10	Lawfulness and compliance	Science-policy link	Dignity

Note: Ethical principles are ranked in each review by the number of separate documents that address a given principle. The definition of specific principles can vary across reviews (e.g., transparency and explainability are grouped together in Jobin et al., but not in the other reviews).

Fairness and justice are consistently at or near the top of these lists of AI ethical principles. Although precise definitions vary across these reviews, fairness and justice typically refers to a need to eliminate, mitigate, or monitor undesirable biases or discrimination. Fairness and justice can be promoted through a variety of means, such as ensuring there are avenues for redress, and by auditing the underlying data, code, and resulting predictions for bias and discrimination.

Transparency can relate to the traceability, openness, and scientific reproducibility of the entire AI pipeline. A transparent process details the data, algorithms, outputs, and resulting actions or decisions. It discloses the extent of human involvement at each step. Further, transparency provides access to the underlying data, code, and other pertinent elements to enable full, independent, scientific review and replicability.

---

[54] 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. See also Hagendorff, Thilo, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines* (2020): 1-22, pg. 112.; Jobin, Anna, Marcello Ienca, and Effy Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389-399.

Interpretability and explainability focus on the extent to which humans can make sense of the model and its outputs. On the surface, this is a matter of understanding the algorithm's nature and how it interacts with a type of data. At a deeper level, it is a desire to understand how an algorithm arrived at a given output, what the key contributing factors were for reaching that output, and whether the output would remain the same if the contributing factors were somehow different.

Accountability refers to the difficult question of determining where responsibility should lie. Numerous often-unrelated individuals and organizations contribute to the development and implementation of an ML or AI system. Data may be collected by one organization, curated by another, and then manipulated by a third to create the actual data that feeds an algorithm. This is based on code likely developed by numerous organizations to produce outputs that are used by yet other organizations, and which in turn may feed further algorithms. Who is to blame if something is amiss? Assessing liability can be challenging when multiple parties are contributing. It may be appropriate to define conditions under which each party must act in order to avoid liability. Other options may be appropriate as well, but each party should probably have at least some stake in ensuring that risks are mitigated. Defining the precise risks is another challenge. AI guidelines focus much more on avoiding intentional or foreseeable harm (through *safety, security, and non-maleficence* guidelines) than on promoting beneficent outcomes for the common good.[55]

Privacy typically refers to maintaining the sensitivity of data collected on individual people. Privacy has a broad scope, stretching from government records, financial records, employment records, and numerous other sources of data about an individual. Social media, personal electronic devices, and website interaction enable data on individuals to be collected seemingly without the individual's knowledge. Even with disclaimers or lengthy legal consent forms, the precise nature of the data that is being collected is often not immediately transparent to the individual.[56] Personal information is needed to facilitate many ML and AI applications. However, privacy concerns arise when personal data are used in ways that are not sanctioned by the individuals that the data capture information on.

Many of these principles are reflected in the set of ethical principles for AI that the DOD adopted in 2020. The DOD principles "build on the U.S. military's existing ethics framework based on the U.S. Constitution, Title 10 of the U.S. Code, Law of War, existing international

---

[55] Anna Jobin, Marcello Ienca, and Effy Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence 1, no. 9 (2019): 394.

[56] In the European Union, the General Data Protection Regulation (2016/679) requires data collectors to be much more explicit than in the United States. Data collected on an individual must be disclosed to that individual upon request, together with information about the use, sharing, and acquisition of the data. The individual can also request that their information be erased. The California Consumer Privacy Act (Cal. Civ. Code § 1798.100), which became effective in 2020, implements similar protections.

treaties and longstanding norms and values."[57] Table 3 presents the DOD AI Ethical Principles, which are meant to be used in both combat and non-combat contexts.

Table 3. DOD AI Ethical Principles

Principle	Description
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."

The DOD AI Ethical Principles partially overlap with the set of top AI ethical principles in Table 2. For instance, the DOD principle of *equitable* aligns somewhat with the common principles of fairness and justice. The DOD principle of *traceable* is likewise related to the common principle of transparency, and the DOD principle of *governable* is similar to the common principle of accountability. However, there are also discrepancies. For example, the common principles of privacy, interpretability, and explainability are omitted from the DOD AI Ethical Principles. They may be implicit in one or more of the DOD's principles (such as the rather vague *responsible*), but they are not explicitly mentioned.

[57] Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

Given the myriad uses of AI within the DOD, its AI ethical principles are necessarily broad. Although frameworks adopted by the DOD and other organizations share many principles, some people have criticized such attempts to develop broadly accepted principles for AI (the deontological approach). Critics claim that "AI ethics initiatives have thus far largely produced vague, high-level principles and value statements that promise to be action-guiding, but in practice provide few specific recommendations."[58] It may be impracticable for diverse AI stakeholders to concur on anything other than general principles that are often not clearly implementable. Principles provide guiding objectives. However, their lack of specificity requires conscious deliberation to consider how they might apply in each case.

The IEEE argues that the rule-based nature of machine learning algorithms allows for the application of a set of rules that represent moral behavior (e.g., by programming an algorithm to avoid revealing sensitive or private information). However, it also acknowledges that these moral rules may be ineffective for directing the model in real-world situations where morals may conflict with each other or some information may be unknown.[59] As Hagendorff (2020) argues, "The generality and superficiality of ethical guidelines in many cases not only prevents actors from bringing their own practice into line with them, but rather encourages the devolution of ethical responsibility to others."[60]

A dissenting view holds that the provision of ethical principles places the onus upon those developing and using AI systems to be explicit about their compliance with these principles. This is analogous to the idea of implementing Institutional Review Boards for research involving human subjects: the ethical principles that need to be followed in research are specified, and the researchers must demonstrate to the review board that their protocols for conducting the research are sufficiently in line with the ethical principles.

Due to the lack of historical norms, collective professional identity, and governance mechanisms in the field of AI, critics worry that ethics guidelines are both unlikely to lead to self-governance among AI practitioners and will likely be difficult to enforce externally.[61]

---

58 Brent Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence* (2019): 501–507, pg. 501.

59 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019, pg. 47, https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

60 Thilo Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines* (2020): 1–22, pg. 112.

61 Brent Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence* (2019): 1–7; Hagendorff, Thilo, "The ethics of Ai ethics: An evaluation of guidelines," *Minds and Machines* (2020): 1–22.

3. Virtue ethics approach to AI ethics

Instead of the primarily deontological approach that AI ethics guidelines represent, others see value in the application of a virtue ethics approach. Virtue ethics—with its focus on cultivating the character traits of virtue, moral wisdom, and eudaimonia—shifts the focus from the ML or AI tools to the developers and users. Instead of attempting to translate universal principles into design requirements, virtue ethics would allow for iteration, learning, and moral wisdom in the application of these principles, with consideration for the context and effects of the algorithm.[62]

This approach has received less attention in the field of AI ethics compared to the deontological approach, but some authors have made suggestions for its application. A caution of solely using the deontological approach is that ethical compliance may become a rote "check-the-box" activity. An inflexible rules-based approach can miss necessary context and may not appropriately empower human decision makers.[63] The development of virtue and moral wisdom in the individual needs to be complemented with study of the ethical issues at play within the spheres of ML and AI. AI ethics training programs and university courses, for example, could facilitate some of the learning and growth necessary for developers and users to meaningfully consider and act upon the ethical implications of their work. Case studies that require stakeholders to work through a diversity of ethical challenges in different contexts could increase understanding of and the ability to negotiate the ethical implications of design choices and applications of AI.[64] This learning and growth process may instill a sense of personal responsibility among AI developers and users for the implications of their work.[65]

Other stakeholders could likewise become motivated to consider the array of short- and long-term effects that their work may have, and to identify alternate approaches that may mitigate any moral and ethical concerns. Hagendorff (2020) argues for a pluralistic approach that incorporates both deontological and virtue ethics. Virtue ethics provides an avenue for infusing the content of ethics guidelines into everyday decision-making by intentionally cultivating moral character and moral wisdom among AI stakeholders. Given the complexity of many AI ethics problems, the differing perspectives of consequentialism, deontology, and virtue ethics are all likely needed to illuminate and evaluate potential solutions.

---

62 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019, pg. 47 https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

63 See, for instance, Sarah Daly, Metin Toksoz-Exley, "Expanding the Ethical AI Conversation: Virtue and its Implications for the Development and Use of Artificial Intelligence Enabled Capabilities [Shortened]," Institute for Defense Analyses, 2022. (Distribution is controlled by DARPA.)

64 Scott Nestler, "Data Ethics and Decision-Making" (presentation, Institute for Defense Analyses, 14 July 2020); Mittelstadt, 2019.

65 Thilo Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines (2020): 1–22, pg. 112..

4. Legal Framework

Laws applicable to personnel policy typically do not differentiate between whether an employment practice is implemented with or without the assistance of technology. Technology is a tool, but the responsibility for maintaining equitable and just employment practices rests with the employer. Employment decisions that unlawful in the absence of machine learning are presumably unlawful in its presence.

However, decisions accepted as lawful in the absence of machine learning may not automatically be viewed as lawful in the presence of machine learning. The involvement of machine learning in a personnel process may invite a higher level of public and legal scrutiny to ensure that the law is indeed being followed. Given the recency of many machine learning advances, specific legal and regulatory requirements regarding appropriate uses of machine learning in personnel management processes are far from being fully delineated. Policies are actively taking shape.

For instance, Executive Order 13960 (Pres. Donald Trump, 3 December 2020) emphasizes both the need for Federal Government agencies to embrace machine learning and artificial intelligence when the "benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed" (a consequentialist approach). However, the Executive Order also cites the need to ensure that use cases exhibit "due respect for our Nation's values" and are "consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties" (a deontologist approach).

The Executive Order also outlines several specific principles that Federal Government agencies must follow in using machine learning and artificial intelligence. These principles include ensuring that applications are consistent with the purpose for which a model was trained; maintaining proper security against "adversarial manipulation" and "malicious exploitation;" ensuring that operations and outcomes are "sufficiently understandable by subject matter experts, users, and others;" documenting the end-to-end pipeline of a model for full traceability; and testing applications regularly, with mechanisms in place "to supersede, disengage, or deactivate existing applications" if problems arise.

Although the legal landscape will continue to evolve, we discuss the general legal and regulatory environment covering personnel management — including core principles that have shaped the existing landscape.[66]

[66] The contents of this section are not legal advice. Qualified legal professionals should be sought for legal advice on specific legal questions.

A. Anti-discrimination Law: General Context

A central legal pillar in personnel management is antidiscrimination law. The Civil Rights Act of 1964 bars employers from discriminating against an individual "because of such individual's race, color, religion, sex, or national origin" (Pub L. 88-352, Sec. 703(a)(1), 2 July 1964). This set of protected classes has since been extended by subsequent legislation and court rulings to include age (for those 40 or older), disability, genetic information, pregnancy, sexual orientation, and gender identity.[67]

In the decades since the passage of the Civil Rights Act of 1964, there has been substantial legal debate about what it means to discriminate against an individual "because of" their membership in a protected class. Discrimination can fall along a broad spectrum from the blatant and malicious to the unintentional. Identifying intent—and proving it in court—can be a challenge. In the 1960s era of broad segregation, an employer might simply impose what may appear to be a neutral policy as a guise for perpetuating discrimination.

Even without any intent to discriminate, an employer could still have policies, procedures, or infrastructure in place that result in unequal outcomes across protected classes.[68] Outcomes are often observable even when discriminatory treatment may not be. However, outcomes do not always reflect discriminatory treatment.[69] A job that routinely requires workers to carry loads of 100 pounds or more is likely going to employ more men than women.

In *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), the Supreme Court ruled that employment practices with a *disparate impact* on protected classes are in violation of the 1964 Civil Rights Act

---

67 "Who is protected from employment discrimination?" U.S. Equal Employment Opportunity Commission, https://www.eeoc.gov/employers/small-business/3-who-protected-employment-discrimination (accessed 19 October 2020). Veterans also have protected status, but it is limited to employers who engage in business with the Federal Government. "Protected Veterans' Rights." U.S. Department of Labor, Office of Federal Contract Compliance Programs, September 2016, https://www.dol.gov/sites/dolgov/files/ofccp/regs/compliance/factsheets/FACT_Veterans_Sept16_ENGESQA508c.pdf. Sundry Federal, State, and local laws prohibit employment discrimination based on other characteristics (e.g., the federal Bankruptcy Reform Act of 1978 prohibits employment discrimination against those who have declared bankruptcy).

68 As an example, even with a policy that permits women to serve as fighter pilots, the vast majority of women in the Air Force have been excluded from piloting an F-15 because it was designed according to a 1967 set of specifications based on the male body size. The Air Force recently mandated that future weapons systems should accommodate a broader range of body sizes. See Valerie Insinna, "To get more female pilots, the Air Force is changing the way it designs weapons," Air Force Times, 19 August 2020, https://www.airforcetimes.com/news/your-air-force/2020/08/19/to-get-more-female-pilots-the-air-force-is-changing-the-way-it-designs-weapons/.

69 Within the military, outcomes that are notably unequal can be counterproductive to morale, cohesion, the ability to execute a mission, and the ability to recruit and retain individuals with broad and diverse skills and experiences. The military's concern for building and maintaining a diverse workforce extends far beyond merely avoiding illegal discrimination. Consequently, machine learning algorithms that merely meet legal requirements will not produce the degree of diversity that the military seeks.

unless those employment practices constitute a *business necessity*. Disparate impact is outcome based. In other words, the Civil Rights Act was not limited to discriminatory treatment but extended to differences in observable outcomes, so long as the observable outcomes could not be justified or explained by the nature of the work. The ability to carry heavy loads may be a justifiable employment requirement for a construction job but not for a desk job.[70] Without the disparate impact criterion, the Civil Rights Act would be difficult to enforce. Its rationale was largely a pragmatic one. According to the unanimous opinion authored by Chief Justice Warren Burger:

> *"The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to […] remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices."*

A subsequent unanimous ruling in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973) established procedures for what evidence individuals and employers needed to demonstrate in order to prevail in a suit. The individual had to first demonstrate prima facie evidence of discrimination. The employer then had the chance to respond by showing that any action taken on their part was non-discriminatory — such as by demonstrating that a disparate impact in the outcomes across protected classes is driven by a business necessity. The individual could then respond in an effort to demonstrate that the employer's actions or policies were really a pretext for underlying discrimination.

Racial or gender disparity in the employer's workforce (as a disparate impact) could be used as a piece of prima facie evidence of discrimination under *Griggs*. However, by *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989), the Supreme Court began to significantly narrow the scope of disparate impact claims, because "any employer having a racially imbalanced segment of its workforce could be haled into court and made to undertake the expensive and time-consuming task of defending the business necessity of its selection methods" (490 U. S. 643).

The 5-4 ruling in *Wards Cove* limited the application of disparate impact enough that Congress voted to overrule it in the Civil Rights Act of 1991 in order "to respond to recent decisions of the Supreme Court" and "to provide adequate protection to victims of discrimination" (Pub. L. 102-166, 21 Nov 1991, Sec 3.2). Another explicit purpose of that Act was "to codify the

---

[70] In *Griggs*, the Duke Power Company had instituted a standardized test for being hired and advancing in the organization. The test was not directly linked to specific tasks of the job, and it disproportionately favored White employees over Black ones. Prior to the Civil Rights Act of 1964, Duke Power Company had openly discriminated against Black employees.

concepts of 'business necessity' and 'job related' enunciated by the Supreme Court in *Griggs*." However, even with this codification, *business necessity* has been a difficult concept for the courts to interpret.

A series of rulings in *Lanning v. Southeastern Pennsylvania Transit Authority (SEPTA)* provides an example of the challenge of identifying a valid business necessity. SEPTA had developed a fitness test used for screening transit police officer applicants as part of a multi-pronged plan to address SEPTA's concerns about public safety. However, the screening policy had a disparate impact against female job applicants. Plaintiffs filed suit under the Civil Rights Act of 1991. The U.S. District Court initially ruled that the fitness test did indeed constitute a business necessity (WL 341605, E.D.Pa. 1998).

On appeal, however, the U.S. Third Circuit Court of Appeals rejected the District Court's ruling, citing that the fitness test did not constitute a *minimum standard* as required by *Griggs*. That is, some transit police officers who were hired before the fitness requirements were implemented would not have passed the fitness test, yet these officers were retained and, in some cases, earned special recognition for their effectiveness. These unfit yet effective officers demonstrated that public safety could be achieved with a lower fitness standard (181 F.3d 478 (3d Cir.1999, *cert. denied* 120 S.Ct. 970 (2000))).

The Third Circuit remanded the case to the District Court to determine whether or not the fitness test was, indeed, a minimum standard. The District Court then determined that SEPTA had met its burden of establishing that the fitness test *did* establish a minimum standard, despite the counter-examples noted above (WL 1790125, E.D.Pa. 2000). The Third Circuit then affirmed the District Court's ruling, with the majority opinion citing a large discrepancy in arrest rates for police officers above the fitness threshold and officers below (308 F.3d 286, 3rd Cir. 2002).[71]

The precise classification of a *minimum standard* provides a useful legal framework for considering an aspect of fairness and equity — concepts that are often included on lists of ethical principles for machine learning. Eligibility cut-offs are often artificial in the sense that some people below the cut-off may, if given the chance, perform satisfactorily. Conversely, some people above the cut-off may fail to perform adequately. The presumption is that those below the cut-off will, on average, have a much lower probability of performing satisfactorily than those above the cut-off. It is a statistical break point rather than one that perfectly discerns satisfactory performers in every single case.

From a statistical standpoint, there are two ways to be right, and two ways to be wrong in establishing a cut-off. A cut-off is "right" to the extent that satisfactory performers are above the cut-off (statistically referred to as true positives), and unsatisfactory performers are below the cut-off (referred to as true negatives). The cut-off is "wrong" to the extent that satisfactory performers

---

[71] The rulings also weighed other factors, such as the severity of the disparate impact, and the observation that SEPTA's fitness standards were higher than those of other law enforcement agencies.

are below the cut-off (referred to as false negatives), or unsatisfactory performers are above the cut-off (referred to as false positives).

A core issue in the SEPTA case is whether the goal is to entirely avoid false negatives. Should the minimum standard be set low enough that no potentially effective police officers are missed (i.e., entirely avoiding false negatives), even if the standard fails to screen out many potentially ineffective officers (i.e., failing to reduce false positives)? Or can the minimum standard be set high enough so that it statistically differentiates a critical mass of potentially effective police officers from potentially ineffective ones? For any given eligibility cut-off, entirely avoiding false negatives is likely untenable — regardless of whether the cut-off is implemented within or outside of an algorithm. The court ruling affirmed this.

SEPTA prevailed in part by demonstrating that those who met the fitness standard had a 70 percent to 90 percent success rate for meeting a set of job standards, while those who did not meet the fitness standard had a success rate on these job standards of only 5 percent to 20 percent. In affirming the District Court's ruling, the Third Circuit reasoned that while a fitness standard corresponding to a predicted success rate of 100 percent would "clearly be unreasonable," demanding a chance of success better than 5 percent to 20 percent "is perfectly reasonable," especially given the public safety implications for having capable police officers (308 F.3d 291, 3rd Cir. 2002).

This type of statistical reasoning is especially pertinent to machine learning applications. A machine learning model may predict the probability that a given outcome will happen. In the military personnel context, this may be the probability that a freshman participating in the Reserve Officer Training Corp (ROTC) will finish college and become a commissioned officer. Or it may be the probability that a new recruit will successfully complete Initial Entry Training, or the probability that a junior officer nearing the end of his or her initial service commitment will continue to serve at a high level of performance for at least five more years.

The DOD may seek to make investments in those who have a higher than average chance of success. Determining who to invest in inevitably involves some form of a cut-off, even if it is a flexible one that considers multiple factors. Of course, depending on the magnitude of the investment and the resources available, the cut-off may be high or low. However, the underlying question remains of how to set a fair threshold that also can statistically differentiate high performers.

Another set of anti-discrimination lawsuits have challenged affirmative action policies as a form of discrimination that disadvantages those in a majority population relative to those in a minority population. The 14th Amendment to the U.S. Constitution includes a provision that each state shall not "deny to any person within its jurisdiction the equal protection of the laws." The emphasis on equality in this clause provides a counterpoint to some affirmative action policies. In *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978), a white male student was

denied acceptance to medical school while minorities with inferior academic credentials were admitted due to a strict quota system that reserved a certain number of seats for minorities.

The Supreme Court's ruling was divided between two plurality opinions: one plurality contended that the use of racial quotas violated the Civil Rights Act of 1964, while the other plurality contended that race could be a factor that universities consider in admission decisions. Justice Lewis Powell offered the deciding opinion, threading the needle between the two plurality opinions. Strict racial quotas imposed by a government school were deemed impermissible.[72] However, race could be used as one of many admission criteria in the spirit of fostering a diverse academic community. Justice Powell's use of diversity as a sufficient justification for considering race was pivotal in changing the dialogue of affirmative action from rectifying social wrongs to valuing diversity for its own sake.

United Steelworkers of Am. v. Weber 443 U.S. 193 (1979) challenged a private employer's affirmative action plan that reserved at least half the spots in a training program for Black employees "until the percentage of Black skilled craft workers in the plant approximated the percentage of Blacks in the local labor force" (443 U.S. 193). Although the ruling expressly did not "define the line of demarcation between permissible and impermissible affirmative action plans," it noted that Congress did not intend to prohibit all forms of race-conscious affirmative action (443 U.S. 195). This plan was permissible because it was temporary, did not unduly hamper the interests and opportunities of white employees, and opened previously unavailable opportunities to Black employees. These and other cases illustrate that affirmative action plans must be carefully balanced. Diversity can be gently encouraged, but not dictated.

The balance between affirmative action and the equality clause of the 14th Amendment is indicative of a broader principle that personnel practices must sometimes satisfy competing objectives. In those instances, the solution requires a thoughtful balancing of the requirements, subject to judicial review. Ethical requirements can likewise have competing objectives, and balance should likewise be sought in such cases.

B. Anti-discrimination Law: Machine Learning Applications

The above legal cases give a taste for how courts have interpreted and enforced anti-discrimination laws. Disparate impact is used as an initial threshold for further investigation in the absence of direct evidence of discriminatory intent. A disparate impact, of itself, does not automatically qualify as discriminatory. In particular, a disparate impact may be justified as non-discriminatory if it is the result of a business necessity, such as needing certain physical characteristics to successfully perform a job. Although the intersection of these principles with

---

[72] *Gratz v. Bollinger*, 539 U.S. 244 (2003) argued whether the University of Michigan had created a system that was effectively like a quota by awarding admission points to underrepresented minorities. The court ruled that automatically giving a minority admission points did not allow for sufficient individual consideration of the merits of each applicant.

machine learning applications is still nascent, we provide a couple of examples to illustrate the types of issues that may arise.

In November 2019, the Apple credit card came under scrutiny when a prominent computer programmer tweeted that he had received 20-times the credit limit as his wife. The cofounder of Apple likewise received a credit limit 10-times the size of his wife's. The limits are determined by an algorithm at Goldman Sachs that synthesizes information on an individual's credit reports, as compiled by the major credit bureaus. An underlying question is whether the algorithm is systematically discriminating against women, or whether there are other meaningful reasons for these discrepancies.

On average, women have more open credit cards than men, higher rates of revolving credit use, higher installment loan balances, and higher rates of bankruptcy than men with comparable characteristics.[73] An algorithm based solely on credit worthiness factors may accordingly result in credit approval rates and credit limits that differ across gender. However, that does not explain the magnitude of the differences above. It could be that the credit reports for these two executives (the cofounder of Apple and the developer of Ruby on Rails) are different enough from the general population that the result was anomalous.[74] However, that begs the question of how anomalous results should be adjudicated. Part of the complaint against Apple and Goldman Sachs was that there was no clear course of action for contesting the results of the algorithm.

As another example, over the last two decades, a growing number of court systems within the United States have incorporated algorithms for predicting recidivism into their criminal justice systems. Such algorithms incorporate information on an individual's criminal record, as well as other information gathered on individuals who are being charged with crimes.[75] Proponents for the use of such algorithms view them as a way to help determine whether an individual would be a good candidate for an alternative treatment or rehabilitation program. Proponents also argue that these algorithms can help judges gauge appropriate probation conditions. However, the actual use of these algorithms in the courtroom has brought multiple issues to light.

First, judges need training on the intended purpose of these algorithmic risk measures. Some judges read more into these risk measures than may be warranted — to the point where the risk measures may prominently factor into a ruling for whether and how long an individual should be sentenced to prison. This presents a potential challenge to due process, particularly since many of

---

[73] Kim Elsesser, "Maybe the Apple and Goldman Sachs Credit Card isn't Gender Biased," Forbes, November 14, 2019, https://www.forbes.com/sites/kimelsesser/2019/11/14/maybe-the-apple-and-goldman-sachs-credit-card-isnt-gender-biased/?sh=13edc47c1518.

[74] The performance of machine learning models declines when there is limited data to train on. It could be that there is not an adequate sample of wealthy executives in the training data.

[75] For instance, one prominent algorithm uses a battery of 137 questions on criminal history, school disciplinary history, peer group, home life stability, family criminal background, and psychological assessments. See https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html.

these algorithms are proprietary and the accused may not be able to challenge the model's underlying data and assumptions. For instance, the defendant may not be able to provide counter arguments for whether the algorithm is relying solely on information admissible in court, or examine whether risk measures produced by the algorithm are unbiased for a given demographic.

Second, models were implemented without explicitly examining the potential for bias across demographic groups. An investigative study of a prominent recidivism algorithm found that Black defendants were much more likely to be misclassified as high risk for re-offending than White defendants (a difference in the false positives), even after controlling for such things as age, gender, and criminal history. To widen the racial gap further, White defendants who later went on to re-offend were more likely to be misclassified as low risk (a difference in the false negatives).[76] This difference emphasizes the need to at least check whether false positive and false negative rates differ across socially salient or legally protected demographic groups. Even if other constraints (such as the technical limitations discussed in Chapter 1) limit the feasibility of attempting to minimize differences across groups in false positive and false negative rates, practitioners who use the results should at least be aware of the differences. Practitioners should also have a sufficient understanding of what the algorithm is doing and the information used in the algorithm in order to exercise an appropriate level of human oversight.

Third, detecting bias in the historic data used to train machine learning models is not straightforward. Individuals in one demographic group may participate in a particular type of crime at a higher rate than individuals in another demographic group. This is referred to as differential participation. Differential participation in crime is not a matter of discrimination. Rather, it reflects the choices of individuals to participate in a crime from one demographic group versus another.[77]

What can be discriminatory is when — after choosing to participate in a crime — individuals in one demographic group are more prone to be apprehended, incarcerated, or harshly sentenced than individuals in another demographic group. This can happen through bias in juries, judges, prosecutors, police, or others involved in the law enforcement system. Bias in a law enforcement system against a demographic group can result in data that disproportionately represents that

---

[76] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," ProPublica, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; and Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. A separate study on another criminal risk assessment tool found that although the tool produced a difference by race, most of the difference could be attributed to criminal history. See Jennifer L. Skeem and Christopher Lowenkamp, "Risk, Race, & Recidivism: Predictive Bias and Disparate Impact," 14 June 2016, http://dx.doi.org/10.2139/ssrn.2687339.

[77] Participation in a crime may be influenced by many factors such as the local prevalence of the crime, social tensions, the probability of apprehension, the expected severity of punishment, and the degree to which the rule of law is enforceable in an area. In some cases, it may be that discriminatory pressures contribute to social tensions, which in turn contribute to participation in a crime. However, the U.S. legal system is premised on individual accountability: the individual still has the choice of whether to participate in a crime or not.

demographic group's actual participation in crimes relative to another group. This is referred to as differential selection. Absent additional circumstantial evidence that may reveal bias in a law enforcement system, disentangling differential participation from differential selection in criminal data can be challenging at best and impossible at worst. In some cases, it simply may not be possible to tell whether differences in crime rates across demographic groups reflect a systemic bias in the law enforcement system, a difference in cultural propensities to participate in a crime, or some combination of each.

These examples illustrate the need to explore and test the robustness of machine learning applications across salient demographic groups. When testing reveals disparate impacts, potential sources of bias should be examined and considered. Identifying, for instance, that data were collected in a way that somehow misrepresents a demographic group is a fundamentally different issue than data that faithfully represent processes that have imbedded social biases. We discuss some possible corrections in Chapter 1. However, corrections are not always available or adequately robust. To the extent that biases may be present, users of the machine learning predictions should be informed of any shortcomings and be trained on appropriate interpretations and uses of the predictions. When individual predictions are challenged, it may also be prudent to have appropriate venues for appeals.

C. Other legal considerations

Antidiscrimination is a prominent legal consideration for the implementation of various machine learning techniques, but it is certainly not the only legal consideration. Under our taxonomy, machine learning transitions to AI when the machine (not the human) makes a decision that involves more algorithmic sophistication than following a set of hard-coded "if-then" classification statements. However, even were the machine learning algorithm merely meant to synthesize data, it may cross the AI threshold of decision making, if the human decision makers either place too much confidence in the algorithm's output or cannot place less emphasis on synthesized data.

Rubber stamping decisions suggested by machine learning algorithms may happen, for instance, due to a limited understanding of an algorithm's caveats or due to limited bandwidth for scrutinizing each individual decision. Rubber stamping may also happen by degrees. For instance, if the machine learning predictions are output as probabilities, a human decision maker may effectively rubber stamp predictions that are above or below some threshold probability, but more carefully assess those in the middle. Although separating straightforward cases from more borderline cases may serve as a good heuristic, checks should be in place to ensure that the straightforward cases are indeed straightforward. Randomly sampling the seemingly straightforward cases for closer review could accomplish this.

Within federal agencies, to the extent that humans defer decisions to algorithms, those algorithms may be interpreted as encoding agencies priorities. If so, they "may qualify as 'rules' under administrative law," and thus require public inspection "via the notice-and-review comment

process or exposing them to pre-enforcement judicial review."[78] Full public review of algorithms and their underlying data may be untenable for a number of reasons. The underlying data may not be releasable due to classification status or other limitations that exempt the data from the Freedom of Information Act (5 U.S.C. § 552). Personnel data, in particular, are protected by the Privacy Act of 1974 (5 U.S.C. § 552a). Algorithmic details may be business proprietary information or otherwise protected by intellectual property laws.

Moreover, even if the algorithms and data are accessible for review, they may be too complex for decision makers and the agencies they represent to invest the time or resources to understand their implications. De facto acceptance of algorithmic outputs may be likely even if the full pipeline of code and data are available for review.

If algorithms or decisions based on them become subject to the Administrative Procedure Act (Pub L. 79-404, 11 June 1946), then the decisions must not be arbitrary or capricious. It is not clear what standards algorithms will be judged by to demonstrate that their output classifications are neither arbitrary nor capricious. On the one hand, algorithms can make the factors leading to a decision more transparent and explicit than many current processes allow. But on the other hand, as discussed in the next chapter, there are mathematical limitations to satisfying multiple ethical objectives in an algorithmic process.

Satisfying one objective may necessarily mean that another cannot be satisfied — even if the decision maker desires to satisfy both. This can force individuals and organizations using these models to rank ethical prerogatives that they would prefer not to rank. In that regard, having such an explicit connection between the inputs and outputs of a decision process may arguably provide too much transparency. However, given the internal complexity of many machine learning models, tracing the connection between the inputs and outputs can still take considerable work.

Another legal question is liability. In a complex system involving both human interactions and sophisticated computer or robotic processes, who is to blame when something goes wrong? Are the engineers liable for failing to provide the human operators with a needed capability? Are the programmers at fault for algorithmic choices that may, under certain conditions, produce problematic results? Who bears the burden of testing the performance of each component, either individually or as an integrated whole? When is a test sufficient? How much burden should be placed on the end user, who may have limited technical understanding of the system but is empowered to act in some way?

From a legal standpoint, the liability burden is not evenly split. Looking at the history of autopilot systems in airplanes as an early example of human interaction with complex computer

---

[78] David F. Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar, "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," report submitted to the Administrative Conference of the United States, February 2020, p. 76, https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

or robotic-like processes, Elish (2020) noted "while automated systems were being relied on more, the nearest human operators were being blamed for the accidents and shortcomings of the purported 'foolproof' technology." Even when the automated system restricted the human operator's ability to act in a critical way, the nearest human operator legally tended to bear the brunt of the entire system's failure.[79] Simply placing a human in the loop is not a guarantee that the system as a whole will work as planned. The system may have an underlying flaw that the human may not be equipped to correct (or equipped to correct within a pertinent window of opportunity). Applications of human-machine teaming need to be examined, tested, and validated.

Another dimension of liability is that many machine learning algorithms incorporate some degree of open-source code. A common element of open-source licenses is a disclaimer whereby the software is provided on an "as is" basis with no warranty.[80] These licenses can shield both the upstream and more proximal software developers from liability. This fits with the pattern from the autopilot systems. For incorporating machine learning into the personnel management context, the largest legal liability for potential failure or misuse may likely be on those who are closest to the actual use of the synthesized information.

This is just a sampling of the broad range of legal considerations that may affect the application of machine learning in the personnel management context.

---

[79] Madeleine C. Elish, "Who is Responsible when Autonomous Systems Fail?" *Center for International Governance Innovation*, June 15, 2020, https://www.cigionline.org/articles/who-responsible-when-autonomous-systems-fail. See also Madeleine C. Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interactions," *Engaging Science, Technology, and Society* 5 (2019): 40–60.

[80] For instance, the MIT License, a fairly permissive open-source license, contains the following in all capital letters (the remainder of the license uses normal capitalization): "THE SOFTWARE IS PROVIDED 'AS IS', WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE." Most other licenses contain similar language. The full text of this and other commonly used open-source software licenses can be found at https://opensource.org/licenses.

This page is intentionally blank.

5. Machine Learning Methods for Equality and Discrimination Concerns

A machine learning model can be viewed as a mirror: it reflects a detailed picture of its input data. If there are biases or blemishes in the data, they will be reflected in the model's output. In a similar vein, data themselves are a mirror of processes in the world. To the extent that data faithfully capture and represent their real-world analogs, injustices that exist in the world will be reflected in the data. Machine learning models can be quick to identify existing social patterns. However, they are agnostic as to whether those patterns represent saintly virtues or social pathologies. Diagnosing and correcting undesirable patterns is left to human hands.

That said, machine learning models can provide clues for where humans might look to investigate nuanced sources of inequality. By looking at how the model itself ascribes predictive importance to data features by demographic group, the model can be used as an explorative tool for identifying correlates that may impact one group differentially from another. The value of the machine learning model in this regard is to sort through vast amounts of data to focus on a smaller set of high-impact correlates. Identifying whether the correlates have any causal meaning naturally requires further investigation. In this regard, the machine learning model provides a learning opportunity to identify differences and explore why they might exist.

Recognizing and learning from differences is one thing. Other applications require more active mitigation. To a limited extent, compensating corrections can be used to help close the gap across some dimensions of inequality. Mathematically, not all dimensions can be closed simultaneously, so the user must focus on a particular dimension. These methods build on a recent literature exploring mathematical implementations of fairness.

This chapter explores these topics to provide a survey of machine learning methods that may be used to identify and, in some cases, compensate for equality concerns.

A. Exploring Differences across Demographic Groups

Just as traditional descriptive statistics and exploratory data analysis can be used to identify patterns within data that may merit further exploration, machine learning models may be used as hypothesis-generating tools. Patterns that the algorithm determines to be pivotal indicators of one outcome compared to another can be probed. These indicators may simply be spurious artifacts of the particular sample that have no real generalizable meaning, or the indicators may be indicative

of a stronger relationship. Further exploration would be required to determine the nature of the relationship — spurious, correlative, or causal.

As a hypothesis-generation tool, machine learning models can be used to compare pivotal indicators across demographic groups. Is the model identifying that the same indicators are pivotal across different demographics, or are there demographic specific indicators that the model is highlighting in making a given prediction?

For example, the Navy has interest in understanding career differences of Naval line officers by race, ethnicity, and gender. To explore an aspect of this issue, IDA researchers examined differences in data features that are predictive of two outcomes in machine learning models: retention and promotion to the rank of O-5 (Commander).[81] The retention model would predict whether an officer would leave the service, and if so, when. The second model would predict if and when an officer would get promoted to O-5, conditional on having served as an officer for at least ten years.

The research found that there are seemingly structural differences in the career progressions across demographic groups, both for retention as officers and reaching the rank of O-5. Researchers investigated this by measuring the effect of each feature on individual predicted outcomes through SHapley Additive exPlanations (SHAP) and comparing the results across demographic groups. The SHAP algorithm calculates how much a feature changes the prediction outcome of the model overall by calculating the difference in prediction probability with and without the feature.[82]

The retention model investigated the features that most changed the probability of an officer staying in service. In this model, *officer primary designator* was more of a pivotal predictor of retention for White males than for several other demographic groups; *assigned unit identification code* was especially important for Black males compared to White males, although it registered as a relatively strong pivotal predictor across demographic groups; and *religious denomination* appeared to matter more for non-White females than for White females. Further differences in retention patterns across demographic groups could also be explored by drilling down into specific values of each pivotal data field. For instance, retention patterns across particular religious denominations or unit identifications codes could be examined by demographic group.

---

[81] Julie Lockwood, Joseph King, and Rachel Augustine, *Explaining Differences in Predicted O-5 Promotion Outcomes by Race and Gender among Naval Officers*, IDA Paper P-20452, (Alexandria, VA: Institute for Defense Analyses, December 2020); and Julie Lockwood, Rachel Augustine, and Joseph King, *Identifying Correlates of Navy Line Officer Retention and Promotion among various Demographic Groups Machine Learning for Hypothesis Generation WEAI 2021*, IDA Paper NS P-22655, (Alexandria, VA: Institute for Defense Analyses, June 2021).

[82] The amount that a feature's removal or addition changes the model's prediction probability depends on the order in which features are removed; therefore, a tree-based method is used for optimization. This allows the algorithm to quantify the features that are most important to the probability generated and gives a more nuanced view of the contribution of any one feature.

The promotion model consisted of all officers who had served at least ten years. *Officer subspecialty* is a strong pivotal predictor for all demographic groups, but the type of subspecialty that predicts promotion varies by demographic group. Again, the model does not show why particular data fields are pivotal within the machine learning model, or why a given data field may be more pivotal for one demographic within the model than another. However, taking a closer look at data fields that the machine learning model views as pivotal can encourage informed questions for further analysis — especially when the pivotal data fields meaningfully differ across demographics.

B. Mathematical Implementations of Fairness

Fairness and disparate impact need to be considered for machine learning models used in decision making. Machine learning can be a powerful tool for synthesizing data and providing detailed and accurate forecasts, but the data fueling these predictions may include previous outcomes that decision makers do not want to perpetuate going forward. An initial question for decision makers is whether to use such data at all. If they want to chart a substantially different course for the future than what occurred in the past, then predictions based on previous outcomes are less relevant; and machine learning is not likely to be an appropriate tool. However, if decision makers are seeking instead to make minor course corrections to what happened in the past, then it may be possible to implement modest compensating corrections within the machine learning context.

For example, if a promotion system were biased, then predictions about who will be promoted would incorporate that bias. Using these predictions in promotion or assignment decisions would, in effect, ratify the existing decision-making system, with all of its biases. Decision makers seeking to remove the biases of such a system would likely need to make meaningful changes to the promotion system itself. Changes to a machine learning model would do little since the promotion system itself needs to be changed. However, there are other cases where a compensating correction to a machine learning model may be appropriate.

The importance of having a model that is "fair" across different groups is based on the type of actions that may be taken in response to the model. If an action has little or no differential impact on specific individuals — such as in the fictional vignette of using a machine learning model to better budget for training needs — then having a model that is "fair" across different demographics is less pertinent. In this case, differential predictions in the model are less likely to result in differential treatment of individuals across groups. For machine learning models that inform more targeted actions — such as in the fictional vignette of targeted retention interventions — it may be more pertinent to assess the ways in which a model may be "fair" or "unfair" across different groups.

A nascent literature on "fairness" in machine learning postulates mathematical criteria for capturing various aspects of what it may mean to be fair (see Mehrabi et al. (2019) and Verma and Rubin (2018) for reviews). Formulae for "fairness" may help provide concrete definitions that

algorithms can use to quickly flag occurrences of "unfairness." However, given that real-world notions of "fairness" are often much more nuanced than these formulae can capture, it is important to remember the human element for interpreting them as one of many potential indicators of "fairness." Too much adherence to these formulae may lead either to a sense of false security in overlooking aspects of fairness that are not formulaically triggered as being unfair, or to flagging things as unfair that do not hold up as such under closer scrutiny.

Another caveat is that it is often mathematically impossible to simultaneously satisfy multiple statistical definitions of fairness (see, for example, Kleinberg et al. 2016). Decision-makers may thus need to consider the merits of each definition and exercise judgement in applying and interpreting fairness tests in order to balance multiple ideals or objectives.

Many fairness criteria focus on certain kinds of disparate impact and may be further limited to specific modeling approaches. For illustrative purposes, we describe a few fairness criteria that are tailored to *binary classification.* A binary classification model predicts whether something will fall into one of two groups, such as whether service members approaching the end of a contract will reenlist. Such a model may correctly predict that a service member will reenlist (a *true positive*), may incorrectly predict that a service member will reenlist (a *false positive*), may correctly predict that a service member will not reenlist (a *true negative*), or may incorrectly predict that a service member will not reenlist (a *false negative*).[83]

For any given classification model, the rates at which a model makes correct predictions (true positives and true negatives) and incorrect predictions (false positives and false negatives) may differ across groups, including protected classes such as race or sex. Three fairness criteria stipulate that they should not, but do so in different ways. The *equal opportunity* fairness criterion stipulates that true positive rates should be identical across relevant groups. The *predictive equality* fairness criterion stipulates that false positive rates should be identical across groups. The *predictive parity* fairness criterion stipulates that precision rates should be equal across groups.[84]

It is mathematically rare and often impossible to satisfy more than one of these criteria at a time. A binary classifier can be tuned to produce a higher true positive rate, but doing so increases the false positive rate. Figure 1 illustrates this tradeoff with *receiver operating characteristics* (ROC) curves for three different groups. ROC curves plot the tradeoff between the true positive

---

[83] In this binary classification example, the service member is predicted to either reenlist or not reenlist. There are only two options. A more general prediction outcome is to predict the probability that the service member will reenlist. If necessary, the probabilities could then be collapsed into two or more groups (e.g., two groups: reenlist vs. not reenlist; three groups: likely to reenlist vs. uncertain vs. likely to not reenlist).

[84] More formally, these criteria require that the following rates are equal across groups.
Equal opportunity: True Positives / (True Positives + False Negatives)
Predictive equality: False Positives / (True Negatives + False Positives)
Predictive parity: True Positives / (True Positives + False Positives)

rate and false positive rate for different classification thresholds.[85] Importantly, ROC curve can often differ across groups, especially when characteristics that are relevant to a classification are not distributed identically across groups. The equal opportunity criterion is satisfied wherever a horizontal line can pass through each of the ROC curves, indicating that the true positive rate is the same across each group.

Likewise, the predictive equality criterion holds whenever a vertical line can pass through each of the ROC curves. Satisfying both the equal opportunity and the predictive equality criteria requires all of the ROC curves to overlap at a particular point. However, there is no guarantee that any two ROC curves will overlap (except at the endpoints of zero and one). If there happens to be one or more serendipitous points where all of the ROC curves overlap, there is no guarantee that the predictive parity criterion will also be simultaneously satisfied.

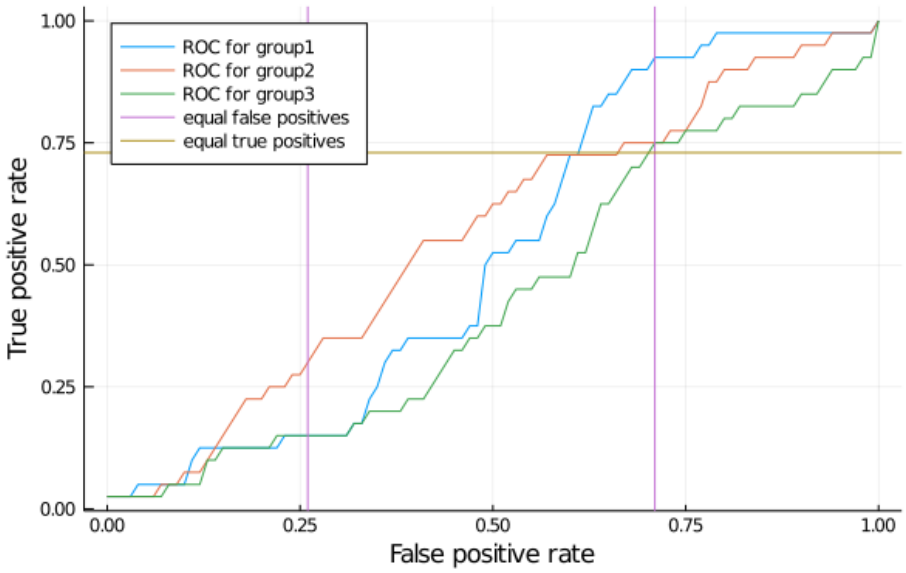

Figure 1. Example of ROC curves for three different groups

A less restrictive test would be to use threshold ranges. That is, the true positive rate is within some threshold range for each group, and the false positive rate is likewise within some threshold range for each group. Using thresholds increases the potential for two or more of the criteria to be simultaneously satisfied. However, there is still no guarantee that this may hold unless the thresholds are sufficiently large.

---

[85] Prior to the final binary classification step, predictions are typically in the form of probabilities. Collapsing probabilities down to a binary outcome requires a classification threshold. Predictions below the threshold are assigned to one of the binary outcomes and predictions above the threshold are assigned to the other outcome. There are numerous potential thresholds. At one extreme, there are no true or false negatives because everything has been classified as a positive. At the other, there are no true or false positives because everything has been classified as a negative. The ROC curve plots the tradeoff between the true positive rate and false positive rate as the threshold systematically varies between these two extremes.

C. Imposing Compensating Corrections

Disparate impact in predictions can be mitigated in one of two general ways. First, the predictors or predictions can be explicitly adjusted on the basis of group membership. For instance, the predictions for one group could be advantaged or disadvantaged in some statistical way relative to those of another group. This, however, would constitute disparate treatment, even if well intentioned. Second, the model itself can be adjusted to push it toward producing more equitable predictions (and away from accurate predictions). This second option may or may not constitute disparate treatment depending on how it is implemented. We present a few examples for reducing disparate impact in predictions. In considering these examples, it is pertinent to keep in mind the legal framework for affirmative action.

Affirmative action seeks to rectify imbalances between groups by explicitly using group membership to affect outcomes. There are multiple mechanisms for accomplishing this. The use of quotas, wherein members of each group compete for pre-allocated slots for that group, is one such mechanism. Another mechanism is to have different standards for different groups. As noted in Chapter 4, the Supreme Court ruled in *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978) that quotas could not be used in the admission process of public universities. However, in *United Steelworkers of Am. v. Weber* 443 U.S. 193 (1979), the Supreme Court ruled that a private employer's affirmative action plan may be permissible under some circumstances. Subsequent rulings have carved out particular forms of affirmative action that are or are not permissible when weighed against the 14th Amendment's equal protection clause and other applicable laws. Given this narrow balance, the corrections described below may only be legal in limited circumstances or may not be legal at all.

1. Pre- and post-processing

Pre-processing techniques alter either data explicitly or alter how data enter into a model. When altered on the basis of group membership, it can change what the model predicts for different groups. Calmon et al. (2017), for instance, recommend replacing individual data with random values drawn from a conditional distribution. The conditional distribution is chosen to minimize group differences in data, but is constrained to avoid extreme changes to the data. In effect, this technique randomly changes data so that those from groups who would be expected to have lower predictions instead have higher predictions and visa-versa. This transformation is used both for model development and training and also on data that enter the model for actual prediction purposes. Since group membership directly determines predictions, decisions based on this technique would constitute disparate treatment.

Post-processing techniques apply transformations to the predictions coming out of a machine learning model. For binary classifiers, Hardt et al. (2016) recommend setting different requirements for different groups so as to satisfy their "equal opportunity" fairness criterion. Specifically, they recommend setting different binary classification thresholds for different groups so that true positive and true negative rates are equivalent for each group. Aligning both the true

positive and the true negative rates is only possible at points where the ROC curves for the different groups intersect; therefore, they also recommend a method for helping to bring the ROC curves into alignment. They add randomness to the predictions for one group but not the other.[86] Adding randomness to one group but not the other effectively degrades the quality of the predictions for that group until it equals the quality of the predictions for the other group.[87] The main difference between this recommendation and traditional affirmative action is that Hardt et al. prioritize equalizing the true positive rates, whereas traditional affirmative action prioritizes equalizing group composition.

2. Penalizing disparate impact

Standard machine learning methods typically maximize one thing: the predictive fit of the model. Machine learning methods can incorporate additional objectives in order to maximize or minimize several things in tandem with predictive fit, such as minimizing disparate impact or maximizing various fairness criteria. Adding additional objectives sacrifices predictive fit. The degree to which predictive fit is sacrificed (and the degree to which other objectives are achieved) depends on how much these other objectives are prioritized (or weighted). Celis et al. (2018) discuss a binary classifier that optimizes over several such additional objectives.[88]

As an informal example, suppose many different performance metrics are used to predict attrition — physical fitness scores, reports of inappropriate behavior, etc. Suppose also that the attrition predictions guide bonus decisions and that women score worse on the fitness tests but exhibit fewer behavioral problems. If, on net, women are predicted to attrite at higher rates, this could lead to a disparate impact on bonuses. A penalty on disparate impact would cause the model to put more emphasis on behavior problems and less on physical fitness. Thus, predictions based on this penalized model would not, in aggregate, favor men over women to the same extent as the original model.

This general technique may achieve meaningful but partial reductions in disparate impact without severely sacrificing predictive power. This is a consequence of the optimization process, where small deviations from the optimum are usually still very close to optimal.[89] Small penalties

---

[86] The binary classification for each group is either based on a fixed threshold (i.e., elements above the threshold are classified as positives and those below the threshold are classified as negatives), or a mixture of two thresholds (i.e., those above the upper threshold are classified as positives, those below the bottom threshold are classified as negatives, and those in the middle are randomly assigned to be either positive or negative with some probability).

[87] Although it equalizes the quality of the predictions at the group level, individuals who had negative random noise added to their prediction may be worse off.

[88] Elisa Celis, Lingxiao Huang, Vijay Keswani, Nisheeth K. Vishnoi, "Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees" (2018), arXiv:1806.06055v3.

[89] Technically, the unconstrained optimum of a continuously differentiable objective function is located at a critical point — where the gradient is zero. Deviating from the optimum harms the objective function in proportion to

may thus generate large reductions in disparate impact without sacrificing much predictive power. Larger penalties, however, may reduce predictive power substantially without generating much additional reduction in disparate impact.

Before using such a penalized model to inform decisions, it would be prudent to do a careful comparison of forecasts from the penalized and unpenalized model to understand how much predictive power is being lost and for whom. Such a comparison should also examine not just aggregate changes, but also changes to person-level predictions.

Depending on how the penalization is implemented, this technique may or may not involve disparate treatment. For example, Lipton et al. (2018) argue that disparate impact penalties can perfectly reproduce any form of disparate treatment, if predictors exist that redundantly encode group membership.[90] When predictors merely correlate with group membership, these disparate impact penalties might be interpreted as using proxy-discrimination to reduce disparate impacts. In the attrition example above, a predictive model that penalizes disparate impact places less weight on physical fitness scores and more weight on behavior problems precisely because these proxy for gender. Lipton et al. argue that this is functionally equivalent to disparate treatment.

3. Removing predictors of group membership

Personnel data inherently capture cultural patterns. Broad patterns for how men and women interact with the labor market will almost inevitably be reflected in an individual organization's personnel records. The same is true along other cultural dimensions: religious, racial, geographic, economic, political, etc. Cultural patterns run deep and affect behavior in intricate ways. The totality of the cultural effects of being a particular gender or being part of a particular religious tradition cannot be represented as a single data element. Consequently, efforts to make a model gender neutral or race neutral cannot declare victory by simply removing information on a person's gender or race. Excluding group membership from a model is typically insufficient to ensure that the model does not statistically discriminate on the basis of group membership.

As an example, in the mid-twentieth century, mortgage underwriters engaged in the practice of *red-lining* by rejecting mortgage applications from U.S. Postal Service ZIP codes where the population was predominantly Black. Although this decision process was superficially race neutral in that it did not explicitly discriminate on the basis of race, its intended effect was to discriminate. The policy statistically discriminated on the basis of ZIP code in order to statistically discriminate on the basis of race. Just as red-lining (intentionally) used ZIP code to proxy discriminate on the

---

the gradient. Thus, the harm to the objective function of a small deviation from the optimum is approximately zero. Further away from the optimum, the gradient may be much larger in magnitude. The harm to the objective function is amplified with larger deviations.

[90] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley, "Does mitigating ML's impact disparity require treatment disparity?," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8136–8146, 2018.

basis of race, decisions based on predictive models have the potential to (perhaps unintentionally) proxy discriminate.

To avoid proxy discrimination on group membership, a predictive model must avoid using any component of the prediction that implicitly operates through predictions of group membership. Calders et al. (2009) recommend a pre-processing technique that mitigates proxy discrimination. Specifically, they recommend reweighting data so that group membership is uncorrelated with what is being predicted. For example, if a model were to predict officer promotion and women are promoted at lower rates than men, then the reweighting scheme would duplicate observations where women are promoted, producing a synthetic data set in which men and women are promoted at the same rates. Training a model on such synthetic data ensures that predictors do not predict the outcome merely because they also predict group membership. Calders et al. demonstrate that their method can substantially reduce disparate impact; however, their methodology is not guaranteed to always reduce disparate impact.

Similarly, IDA has produced a methodology that mitigates proxy discrimination by including group membership when fitting a model, then removing group membership from the fitted model before making predictions. Intuitively, when a model excludes group membership, variables predict an outcome partly through predictions of group membership (an indirect effect) and partly through a direct effect. Including group membership when the model is trained ensures that other variables predict the outcome only through their direct effect. After the model is trained, group membership is removed by "integrating out" group membership using rates of group membership, effectively averaging predictions for each individual as though he or she were a composite of all groups. This methodology operates similarly to Calders et. al, but can be applied to arbitrary groups after the model has already been fitted. Furthermore, it allows investigation of the role proxy discrimination may be playing in predictive models.

This page is intentionally blank.

6. Operationalization

Thus far, we have laid out legal, moral, and ethical frameworks for assessing applications of machine learning in the personnel context. This chapter synthesizes the key points of these frameworks, discusses key considerations for legal, moral, and ethical use of machine learning across the project life-cycle, and applies the DOD's Ethical Principles for Artificial Intelligence to the illustrative vignettes from Chapter 2 for using machine learning in personnel management.

A. Summary of Legal, Moral, and Ethical Frameworks

1. Legal

To an extent, laws reflect the ethical determinations of society at a given point in time. American society has determined some ethical principles to be inviolate rights or obligations. The values of equality before the law, equality of opportunity, and consent of the governed are enshrined in the U.S. Constitution and Bill of Rights. Subsequent acts of Congress, federal court rulings, and Executive actions have specified these concepts. For personnel management in particular, anti-discrimination law — particularly the 14th Amendment, the Civil Rights Act of 1964, related legislation, and a number of court rulings — provides a central framework. Such laws govern employment practices regardless of whether these practices use machine learning or other algorithmic techniques. The involvement of machine learning, however, may invite a higher level of scrutiny or the development of more targeted laws and regulations in the future.

The Civil Rights Act of 1964 prohibits employers from discriminating on the basis of race, color, religion, sex, or national origin. The Supreme Court has interpreted this Act to prohibit practices that have a *disparate impact* on individuals in any particular demographic group, unless those practices constitute a *business necessity*. In later rulings the Court clarified the concept of business necessity with the concept of a *minimum standard*, or the minimum qualification necessary to do a job. Discrimination may be a business necessity, if it results from the application of a minimum standard.[91]

[91] For instance, a job that requires regular heavy lifting may screen potential employees for their ability to lift 100-pound loads — a requirement that will, on average, be much harder for females to meet. The standard, however, must be directly connected to the job requirements, and court cases can focus on the degree to which the standard is connected to the job requirements.

Practices that seek to promote diversity may also run afoul of antidiscrimination law. In particular, affirmative action practices have faced challenges as a form of discrimination that disadvantages majority populations. Rulings on affirmative action policies generally strike down strict hiring or admissions quotas but are more permissive of policies that promote diversity in less restrictive ways.

Antidiscrimination law typically considers the individual characteristics specified in the Civil Rights Act of 1964, but these legally protected classes are not the only groups that the DOD may need to avoid inadvertently singling out in a negative way. For instance, personnel policies may need to be cognizant of disparate impacts on service members based on officer, warrant, or enlisted status; officer accession source; first-generation military service vs. multi-generational military service; etc.

2. Moral and Ethical

Drawing from theories of normative ethics, which define standards for moral behavior by classifying phenomena as "right" or "wrong," our application of ethical principles applies multiple schools of thought. The three major schools — Consequentialism, Deontology, and Virtue Ethics — offer contrasting viewpoints on how to approach moral determinations, and certain aspects of each intersect with essential American principles and military values. It is these intersections that we find most applicable to the military personnel context.

Consequentialism, which judges an action's morality by weighing its costs and benefits, is a foundation of many decision-making processes and is broadly applicable to the use of machine learning in DOD personnel processes. In the DOD context, decision makers consider the costs and benefits that individuals, organizations, and the DOD mission as a whole will incur as a result of a policy change. The nature of this cost-benefit analysis may vary. For example, it may focus narrowly on impacts to military readiness or, consistent with a classical utilitarian approach, it may include societal, personal, financial, or equality implications over the short- or long-term.

A practical application of Consequentialism in machine learning may involve training a model to appropriately value desirable and undesirable effects and to maximize good outcomes on average. This valuation may be relatively concrete, such as a decision's financial cost, or more abstract, such as the value of individual rights or equity. The caveat here is that the process of defining, measuring, and balancing desirable and undesirable consequences is often not straightforward and likely entails ethical decisions in and of itself.

Deontology addresses individual rights more directly. These theories hold that all individuals have a duty to act in accordance with pre-defined moral principles and that they have rights upon which others may not infringe. In the contractarian view, individuals in a society accept these principles as part of a social contract that balances the interests of all members of society. In American society, these deontological concepts are manifested in Jeffersonian values of equality of opportunity, equality before the law, and consent of the governed. These values are reflected in

the AI ethics frameworks developed by the DOD, the IEEE, and by many industry stakeholders, to include principles of fairness, justice, beneficence, and transparency.[92]

Using the DOD's AI Ethical Principles to guide the use of machine learning in personnel policy is an inherently deontological approach; this chapter will explore applications of these principles in detail.[93] A broader understanding of deontological ethics may be useful for guiding the operationalization of the DOD's AI Ethical Principles in the personnel management context. For example, the idea of the social contract could inform the application of the DOD's AI Ethical Principles of being Responsible and Equitable. When joining the U.S. military's all-volunteer force, service members have a duty to uphold their oath to defend the nation, but they may also expect DOD personnel policies to reasonably balance this obligation with service member self-interests (such as individual freedom, personal wellbeing, dignity, and professional growth).

Beyond service members' oath to defend the nation, military creeds espouse core values and the development of virtuous character traits. This is in line with Virtue Ethics, the third school of thought in normative ethics, which conceptualizes morality in terms of a person's intrinsic values, moral wisdom in the practical application of those values, and pursuit of the greatest good. In this view, morality is an ongoing process, developed over time through deliberate effort. The DOD applies a virtue ethics lens in personnel policy when it bases promotion decisions in part on assessments of service members' character, leadership, and military comportment.[94] When incorporating machine learning into personnel policy, algorithms must not overlook these essential characteristics. The DOD also applies virtue ethics in its expectation that the individuals making personnel decisions will do so in a fair and unbiased manner; although humans may not be able to do so perfectly, they can strive to improve and implement safeguards to identify weaknesses. Machine learning algorithms can and should be held to the same standard.[95]

---

[92] See the citations in Chapter 3 for examples. For the DOD and IEEE ethical frameworks for AI, see Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, IEEE, 2019.

[93] Although primarily deontological in nature, an example of a more consequentialist perspective of the DOD's AI Ethical Principles is that DOD's international alliances may be strengthened by adherence to common international expectations (at least as shared by our close allies) for how to use AI. In that regard, having and following such a set of principles may be reflective of the costs and benefits of strengthening international alliances.

[94] Since these traits provide value to the military as a whole, they could also enter positively into a utilitarian calculation.

[95] As noted in Chapter 1, it can be mathematically impossible to simultaneously satisfy multiple fairness criteria. Striving to implement a "fair" algorithm may begin with an awareness of what the decision implications may be from using different measures of fairness. It may likewise entail a desire to be conscientious about safeguards and responsible practices to implement throughout the full life-cycle of the machine learning application. The next section offers related suggestions.

B. Life-Cycle Considerations for Machine Learning

A key consideration in introducing machine learning into personnel management practices is to compare current practices with how they might perform with machine learning. Executive Order 13960 (Pres. Donald Trump, 3 December 2020) stipulates that Federal Government agencies should use machine learning and artificial intelligence when the "benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed." Because the ethical and legal risks of incorporating machine learning into personnel policy are sometimes similar to those of implementing personnel policies without machine learning, this raises the question of relative risk.

In the context of the legal, moral, and ethical frameworks described in this paper, can the use of a machine learning application in a given decision-making process maintain or exceed the ethical status quo of the current personnel policy? Can it move personnel policy closer to the ethical ideal? What is the potential harm if this machine learning application is misapplied in the decision-making process, particularly in ways that could reduce the ethical status quo? These questions likely have complex answers, necessitating a closer look at what is gained and what is lost.

Legal, moral, and ethical risks and benefits will depend significantly upon how stakeholders ultimately implement machine learning, not just on how governing bodies write it into policy. Much depends on the underlying design of the machine learning, to include such things as having appropriate data inputs and transparent processes; its rollout and execution, such as how the use of the algorithm is communicated to various audiences; and measures for identifying and assessing risks, together with mitigation plans and remedies for when potential failures happen. As such, when applications of machine learning fall short of the status quo, it is critical to understand why and work toward improvement.[96]

In order to realize any benefits of using machine learning in personnel management, stakeholders must ensure that machine learning is the right tool for the job, the data the model uses are appropriate, the model was developed correctly, and that it is implemented and maintained correctly and safely. Moreover, ongoing evaluations will be needed to respond to evolving conditions, such as changes in ethical expectations, changes to the input data, or changes to the populations to which the results are applied. Table 4 outlines key considerations for all stages of the project life-cycle, from planning to data selection to design to implementation.

[96] The status quo itself need not remain static. Improvements should be welcomed and sought for, whether or not machine learning applications are involved.

Table 4. Life-Cycle Considerations for Machine Learning Projects

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

- What is the ultimate goal? If there are multiple goals, what are the trade-offs? Will machine learning provide appropriate information for achieving these?
- Would a different approach be more effective than a machine learning prediction?
- How would the risks and benefits of using machine learning differ from the status quo (including development risks, financial risks, legal and ethical risks, etc.)?

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

- Why were the data collected? How reliable are the data? Can their provenance be determined? Were there limits imposed on their use at the time of collection?
- How were the data prepared for analysis? What procedures are in place to ensure the data are maintained with the highest level of accuracy?
- Are there processes to periodically check for unintended bias within the data?
- Are there data elements or correlates of data elements that may be impermissible to consider for particular decisions? How will such data elements be dealt with?
- What safeguards are in place to respect privacy and protect the data?

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

- Have the developers consulted with subject matter experts, stakeholders, and groups that will be affected about potential legal, moral, or ethical issues that may arise?
- What dimensions of diversity should developers consider in designing the model?
- Is there end-to-end transparency in the machine learning project, from data collection and acquisition, to data preparation, to model coding, refinement, and testing?
- How will the system be tested and monitored? Is there a robust code review process? Is the pipeline fully reproducible? What tests will be done to monitor the adequacy of results (e.g., using a set of test cases that should be classified in a given way)?
- Is the pipeline sufficiently modular so that if a component is later found to be problematic, it can be readily swapped out?
- Are there corrective actions that should be taken to minimize differential outcomes across given populations? How will corrective actions be evaluated and tested?

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

- Do stakeholders and users understand the appropriate uses and limitations of the machine learning model? Is there effective documentation and training?
- What is the plan for monitoring and evaluating use of the model?
- Are safeguards in place to identify and intervene in case of unintended consequences?

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

Machine learning is not an appropriate tool for all use cases. Its forte is in generating accurate predictions based on historical patterns in a broad corpus of data. But if the data are not representative of the conditions under which predictions would be useful, or if the data are too limited in terms of their accuracy or detail, then a different tool is needed. In some cases, it may be more effective to acquire the necessary information by collecting additional or higher quality data, rather than by attempting to predict the same information with machine learning using less than adequate data.

Machine learning is also adept at identifying clusters of observations with common characteristics. Clustering observations, and identifying prevalent patterns at the group level is, however, often a much easier exercise than determining why a particular observation was clustered in a given group or whether that observation will follow the modal patterns of the group. Group-level forecasts and classifications, in that regard, are likely to be sounder than at the individual level. Consequently, the use of machine learning to inform individualized decisions will likely require additional care and scrutiny that using it to analyze group-level patterns or inform group-level decisions.

Machine learning can also be used in causal analyses. For instance, it can be used as a dimensional reduction technique when there are a large number of features — potentially with complex and non-linear interactions — that need to be controlled for in a causal model.[97] Here, machine learning is not providing the cause and effect relationships, but is rather playing a supporting role in an analysis where the causal relationships are established by other conditions. If causality needs to be established, other tools, such as controlled trials, are appropriate. Stakeholders must consider the information needed to achieve any particular goal and determine whether machine learning is appropriate and feasible.

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

Appropriate data are critical to the successful use of machine learning. To determine whether a dataset is appropriate, developers should consider why and how the data was initially collected. In particular, data may have been collected in a way that introduced bias. This could happen in a number of ways. Certain populations may self-select into providing information about themselves, or they may have been excluded or underrepresented by data collection policies. It is important to be able to identify whether a group is underrepresented, overrepresented, or misrepresented due to how the data were collected. It is likewise important to distinguish between data collection processes and societal or policy factors that may result in a group being underrepresented. For

---

[97] For a survey on how machine learning can contribute to causal analyses in personnel management, see Julie Lockwood, Alan Gelder, Matthew Goldberg, et al., *Leveraging Machine Learning in Defense Analyses*, IDA Paper P-13174, (Alexandria, VA: Institute for Defense Analyses, May 2020).

instance, are women underrepresented in data about individuals in a given occupation because fewer women were sampled or because few women are actually in the occupation?

In addition to the representativeness of the data, developers must determine the quality of the data before using it in their model. Data quality issues are more likely to arise for types of data where the incentives to collect or verify the accuracy of the data are low. For instance, there are natural checks on administrative data on an individual's pay check: individuals are likely to complain if their pay is lower than expected. Financial audits also provide an incentive for ensuring the accuracy of financial data. Conversely, information that is quickly entered may be entered incorrectly, and information that is not reviewed can become outdated.

Developers can assess whether there any procedural checks or incentives in place to ensure that data are recorded correctly and kept up to date. It is also pertinent to consider under what circumstances the data are changed, how the changes are logged, and whether the custodial processes for maintaining the data are sufficient for the data to maintain their integrity. In some cases, data quality may be unknown for a variety of reasons. Perhaps the data have only been used for administrative purposes and have never been explored for analytical purposes. The development of ML algorithms – especially as they are applied to unexplored data – may highlight data quality issues. Understanding these issues can lead in turn to improvements in data quality through better collection, curation, or other data management processes.

Developers must also be cognizant of any limitations placed on subsequent uses of a dataset, such as privacy laws, data use agreements, informed consent restrictions on data usage, human subjects research restrictions, etc. Regardless of the presence of existing data privacy requirements, developers should determine what risks this new use of the data may pose and identify ways to protect privacy and limit unintended uses. For example, developers may need to create safeguards to prevent any output from being reverse engineered in such a way that compromises privacy. They should also consider whether particularly sensitive fields are needed within the analysis itself, or whether they can be obfuscated or entirely removed from the analysis to reduce risk.

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

When designing machine learning models, developers should consult with subject matter experts and stakeholders to understand the legal, moral, and ethical issues that may arise in the development and use of the model. There may be legal constraints on what a system can even be designed to do that need to be understood upfront. Beyond legal concerns related to protected classes, there may also be concerns about other non-protected classes — as is the case in the DOD. A military that reflects the breadth of society in all of its elements (and not just those that are legally protected) is likely to be more capable, reflect a greater diversity of thought, foster closer connections to civilian society, and benefit from better recruiting. Consulting with key stakeholders in a variety of formats (such as interviews, red-teaming exercises, or structured interviews) can bring to light specific factors that should be considered. This is critical to upholding the DOD ethical principles of responsibility and equity.

Traceability is another important concern during model development. This begins with the data to be used in model development. To the extent possible, developers and other stakeholders should determine whether the data are reputable. If data are being collected, there must be a well-documented process for each aspect of the collection process. There should be a clear line of custody for the data from the time they are collected to the time they are used. Any data preparation steps should likewise be documented and be reproducible. These are common principles of good data stewardship, independent of the method that may be used to analyze the data.

For machine learning, in particular, adherence to sound principles is needed to bolster public trust and ensure scientific rigor.[98] The model development process should aim to provide a level of traceability, which can facilitate auditing and quality improvements. To achieve this traceability, model code must be version controlled. Stakeholders should ascertain whether model runs contain metadata on the data used, model code version, hyperparameters, and other information needed to reproduce the model run.

Stakeholders should also be able to understand how and why the model arrives at its outputs. The degree to which stakeholders can understand the model will likely impact how it is used. An opaque model may inadvertently become a model with a high level of autonomy if stakeholders both trust the output and also do not understand why a given output may occur. The intended amount of autonomy between the model and its users should ideally be specified early on.

Technical review can help to ensure both the robustness of the machine learning model and reveal any previously overlooked legal, moral, and ethical concerns. Review processes should be in place for stakeholders to view the entire design pipeline — from a high-level methodological overview down to a fine-grained examination of the code. Such review processes should help ensure the robustness, reproducibility, and validity of the pipeline.

Review processes should also account for an examination of model performance along any demographic dimensions that are likely to be sensitive. Test data sets may need to be developed that capture critical demographic dimensions in order to help illuminate any potential issues.

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

Developing a robust machine learning model does not guarantee its successful use in practice. Those involved in the implementation and use of a model should receive clear guidance on appropriate uses of the information the model produces, together with guidance on the model's limitations and any risks involved in its use. Given that implementation plans may change, it can be important to maintain notional and evolving plans for implementation throughout the development stage. This can help the developers and the users adjust as necessary so that the final product remains compatible with the final implementation plan.

---

[98] For example, Executive Order 13960 states that "the ongoing adoption and acceptance of AI will depend significantly on public trust."

Stakeholders should collaborate with developers to create a monitoring and evaluation plan to assess whether the model is being used appropriately and effectively. This plan should assess if the intended goals of using the model have been successful and other unanticipated effects, which may be more distantly related to the use of the model. For instance, as time progresses, the circumstances under the model's use may begin to differ from the circumstances under which the training data were generated. The plan should assess how models will be updated. Based on the model's intended use and its known limitations, stakeholders can implement safeguards to assist with the identification of unintended consequences and have plans in place for prompt mitigation should they arise.

C. Analysis of Fictional Vignettes in Chapter 2

Chapter 2 provided four vignettes of hypothetical scenarios in which machine learning could contribute to personnel policy. In this section, we apply the DOD's AI Ethical Principles to each of these vignettes. In doing so, we highlight potential legal, moral, and ethical risks and benefits of each use case. This discussion is not intended to determine whether these scenarios constitute legal and ethical use of machine learning, nor do we address every aspect of these scenarios. The discussion herein is rather meant to illustrate how stakeholders might approach similar deliberations with an eye toward promoting the ethical use of machine learning.

1. Equitable

DOD's AI Ethical Principles states that "the Department will take deliberate steps to minimize unintended bias in AI capabilities." As discussed in Chapter 1, machine learning predictions in the personnel context may reflect various forms of bias. To the extent that biases and inequities exist in our broader society, data that faithfully reflect our society will capture those biases. Any biases in the data will in turn be mirrored in machine learning models (and any other data driven model for that matter). In that respect, some degree of bias is unavoidable. That said, "deliberate steps" can still be taken "to minimize unintended bias." Being cognizant of potential sources of bias is one key. Proactively examining the degree to which it is present is another. Such diagnoses allow stakeholders to conceptualize and frame the extent of the problem and to take a measured approach for any desired correction.

a. Focusing on Standards

Machine learning excels in finding similarities. Many algorithms are based on the concept of using observations with similar characteristics and outcomes to create a framework for predicting like outcomes for other observations with similar characteristics. However, machine learning algorithms are blind as to whether the similarities are appropriate or relevant to consider. In contrast, there are often legal and ethical constraints in personnel management settings that determine what similarities or characteristics may or may not be appropriate to consider. For hiring, promotion, assignment selection, and other significant career decisions, an individual's performance and skills are preeminent. Evaluating individuals on performance-based standards is

consistent with the American ethos of equality of opportunity — allowing all, including the underdog, to compete on merit.

Even with a desire to focus on performance and skills, there are challenges, whether or not machine learning is used in a screening process. Performance can be challenging to measure. Defining aspects of performance that are meaningful criteria to screen is likewise an imperfect process. Screening criteria can be too rigid, excluding talented and competent individuals who would be well-suited for a position but who do not meet some aspect of the criteria.[99] Screening criteria can be misaligned with the totality of job performance requirements, perhaps emphasizing metrics that are easily measurable but not representative of work in the position. Screening criteria may also focus on indirect proxies of desired traits. Screening criteria may also become entrenched in tradition and institutional momentum, making it difficult to correct any deficiencies or enable any alternatives.

When screening criteria are overly rigid or misaligned, introducing machine learning into a screening process will not improve these deficiencies. If anything, because machine learning is guided by past observations, it will reinforce these deficiencies. Machine learning also might introduce new screening criteria that may be irrelevant.

As an example, individuals growing up in a particular part of the country may have a strong record of success in a prestigious military program. A machine learning algorithm may hone in on that and assign individuals from that part of the country a stronger score. The success of individuals from this part of the country may simply be happenstance. Alternatively, it may reflect pertinent experiences, skills, or traits that are common in that location. If so, then growing up in that location serves as a proxy for the experiences, skills, or traits that are valuable for succeeding in the prestigious program. Not all individuals from that location may have those attributes, and individuals from other locations do not necessarily lack them. If those experiences, skills, or traits are indeed desirable things to screen on, it would be ideal to screen directly on them rather than indirectly on where they are geographically concentrated.

Indirect screening processes can preclude individuals from competing on merit by shifting the focus away from an individual's skills and toward group-level characteristics. Exceptional performers from groups that tend to have weaker characteristics may be passed by, and mediocre performers from groups that tend to have stronger characteristics may make the cut.

There is also the question of whether performance standards should be tailored to particular groups. For instance, there is ongoing debate as to whether men and women should be held to the same military fitness standards.[100] If a military task requires a minimum level of strength to

---

[99] For instance, some occupational licensing requirements can be overly prescriptive in specifying a precise course of required training. This excludes those who can demonstrate competency in the occupation in some other way.

[100] The Army Combat Fitness Test recently emphasized this debate. The new fitness test initially planned to have gender-neutral standards, but that decision received pushback from members of Congress, saying it would

perform, and each individual will need to perform the task to succeed in a position, then a common standard may be appropriate. However, if the individual needs to have a given level of endurance and fitness, then gender specific standards may be appropriate. The same holds for age-specific fitness standards.

In the fictional vignettes, there are multiple places where reflection on performance standards is warranted. In the promotion board example, there is the question of measuring individual performance to date compared to projected performance. Decisions about future career growth necessarily need to consider projected performance. The individual may have performed well in one role, but without some kind of projection, it is difficult to assess how the individual will do once promoted in another role.

However, that begs the question of how to make a projection that is equitable. Human heuristics, as well as machine learning, may both lead to inappropriate conclusions — such as in the example above that emphasized growing up in a particular part of the country over more direct measures of the desired traits. It may be that data on the desired traits do not readily exist. If so, it may be more appropriate to devote resources to gathering data on the appropriate traits than to try to approximate indirect correlates of the desired traits.

For instance, data on projected performance could be collected through role-playing scenarios of a future role that the individual may have. Such an activity would allow the individual to be evaluated on merit rather than on indirect correlates of success (whether based on heuristic, machine learning, or some other method).

The question of how much weight to place on demonstrated performance versus projected performance surfaces again in the fictional vignette on recruiting and selection for Special Forces. Correlates that are predictive of long, successful careers can range from things that are idiosyncratic to previous cohorts to things that are meaningful to screen on. To identify meaningful screening criteria, machine learning may initially be used as a tool for identifying potentially pertinent correlates.

Further analysis is then necessary to evaluate why these correlates may be pertinent for predicting long, successful careers. This includes investigating how these correlates may interact with current and past personnel policies, how directly these correlates measure performance

---

disadvantage women. Others, including the first woman to become an Army infantry officer (and one of the first two women to graduate from Ranger School), have argued that, "To not require women to meet equal standards in combat arms will not only undermine their credibility, but also place those women, their teammates, and the mission at risk." Ultimately, the fitness test adopted gender and age specific standards. See Kristen Griest, "With Equal Opportunity Comes Equal Responsibility: Lowering Fitness Standards to Accommodate Women will Hurt the Army — And Women," *Modern War Institute at West Point*, 25 February 2021, https://mwi.usma.edu/with-equal-opportunity-comes-equal-responsibility-lowering-fitness-standards-to-accommodate-women-will-hurt-the-army-and-women/. See also Missy Ryan, "Senators Urge Pentagon to Suspend Implementation of Army's New Fitness Test," *Washington Post*, 20 October 2020, https://www.washingtonpost.com/national-security/army-new-fitness-test/2020/10/20/d46660bc-12da-11eb-82af-864652063d61_story.html. Fitness standards by age and gender are available at https://www.armycombatfitnesstest.com/scoringstandards.

characteristics of interest, and whether there are more direct measures of characteristics of interest. In that sense, the machine learning model is the start of a broader investigation rather than an end. Such an investigation may illuminate characteristics that may be used as screening criteria directly, or they may feed back into some further model.

b. Monitoring Eligibility Changes for Demographic Groups

One form of bias that stakeholders may need to be especially cognizant of is the time lag between when a policy is first implemented and when its effects become mainstream. For instance, until recently, women were not allowed to serve in some combat positions. Even after the policy changed, it can take years for the culture of those positions to adapt to the point where it is not an anomaly for women to fill them. It can likewise take years for a critical mass of women to not just enter those positions but rise in the leadership ranks.

The lack of an established history of women filling these positions with any longevity may enter machine learning forecasts of retention as an indication that women currently serving in these positions may exit soon. Machine learning models excel in predicting events that have been observed previously. However, when the event is new or anomalous (such as women entering career paths they were previously excluded from), machine learning predictions should be evaluated cautiously and used in the context of the recent policy change and the evolving cultural conditions. Methodologies for removing predictors of group membership, such as those discussed in Chapter 5, may be helpful in contexts like these.

This insight applies to both the prediction algorithms in the Targeted Army Retention Intervention vignette and the Recruiting and Selecting for Special Forces vignette because each includes career options that until recently were closed to women.[101] Policy changes that have expanded or limited a particular population's ability to serve in a career path should be cataloged so that the machine learning predictions can be viewed more holistically in light of the changes. Some changes may be more dramatic while others are subtle.[102] In the case of the combat roles

---

[101] Although MARSOC is now open to female service members, no females to date have made it through all phases of the training. Sgt. Bailey Weis became the first and to date only female to complete the two phases of MARSOC's Assessment and Selection course in late 2018. However, she was not selected for the subsequent 9-month Individual Training Course training, and therefore was not able to join MARSOC. The Army graduated its first female Green Beret since the gender barrier was lifted in 2020. See Shawn Snow, "First female completes second phase of Marine Raider selection," *MarineTimes*, 22 October 2018, https://www.marinecorpstimes.com/news/your-marine-corps/2018/10/22/first-female-completes-second-phase-of-marine-raider-selection/. See also Kyle Rempfer, "A woman became a Green Beret Thursday, a huge milestone for the Army and the military, but she isn't the first female to earn the title," *ArmyTimes*, 9 July 2020, https://www.armytimes.com/news/your-army/2020/07/09/a-woman-became-a-green-beret-today-a-huge-milestone-for-the-army-and-the-military-but-she-isnt-the-first-female-to-earn-the-title/.

[102] Some policy changes may be more visible because they pertain to larger populations or otherwise receive more attention. However, changes pertaining to smaller populations should not be overlooked. One example (as noted in Chapter 3) is the recent change to the dress and grooming standards that now permits Sikhs to serve while maintaining their religious standards for wearing beards and turbans.

that more recently opened to women, it would be pertinent to document the specific roles affected, how and when the policy was enacted, and any richer context on how the policy has been implemented.

For instance, are women simply permitted to serve in the roles, or have there been concerted efforts to welcome or encourage women into these roles? This information can be coupled with descriptions on the degree to which women are entering these new roles (e.g., how many entered these roles at the time of the policy change, and what has been the trend for women entering and staying in these roles since, and how well have they performed?). Capturing this policy information can be used to both flag personnel populations that may not have established long-term patterns of service in given roles and to add qualitative context for the nature of the change and the short-term responses to the change. Machine learning provides one form of information synthesis; recent changes to the policy context that may be omitted by machine learning algorithms are likewise critical to synthesize and consider.

It can be difficult to gauge when a policy change has matured to the point that its effects are becoming part of the mainstream (especially in cases where the mainstream may itself be in flux or experiencing significant social changes). As populations that were restricted from certain roles begin to have a critical mass, it may be informative to examine the extent to which machine learning predictions for that population would have changed at earlier dates.

One such test is as follows: train multiple versions of a machine learning model for a personnel application but vary the date range of the input data. One model may only consider personnel data up through five years ago. Another model may only consider data up through two years ago. A third may consider personnel data up through the present. The models are otherwise identical except for the period of data over which they are trained. Each model could be used to make predictions based on the *current* information for individuals in the population of interest.

This exercise would illuminate how including more recent time periods — and any associated cultural changes that occurred during those time periods — impacts the predictions. Would an individual, with his or her observed characteristics today, have had the same prediction five years ago or two years ago? If the predictions shift, how do they shift? Do they shift in a similar way for individuals within the population of interest? Is this different from how predictions do or do not shift for individuals in other populations? This exercise may provide useful insights for monitoring populations that have experienced policy changes that have either expanded or limited their access to given roles. Policy changes may drastically and immediately shift social patterns, or the change may take place over time, or they may have negligible effects. It is prudent to record policy changes so that they can be explicitly examined and incorporated into machine learning models.

c. Officer Promotions

To facilitate equitable decision making in the promotions process, promotion review boards operate on the expectation that all members will act without prejudice or partiality. Policies

stipulate measures to minimize the chance of bias, including limits on information the board may consider, to include some demographics. Promotion boards also typically have processes in place to help standardize evaluations. For instance, boards may score a handful of promotion packets from a prior year (for whom promotion decisions have already been made) to enable the board to gain practice and calibrate adjudication standards. To help ensure consistency across the board in the evaluation process, boards may also have rules that trigger a further assessment of a candidate if the spread of scores that a candidate receives is too large. This provides a safeguard against an isolated board member who may be overly optimistic or pessimistic about a candidate.

However, even with these and other safeguards in place, evaluation criteria may unintentionally drift over time; members may overlook or misunderstand certain information; and members may inadvertently bring some bias into their decision-making. Board members may also have a difficult time assessing candidates' military experience that is out of their area of expertise (such as Joint assignments that board members may be less familiar with).

Machine learning can provide a check on this, particularly if members' assessments differ substantially from the model's predictions. In this regard, machine learning could hypothetically be used at multiple steps to mitigate bias. For instance, an addendum to this vignette might be:

> *Ultimately the board members decide which officers to recommend for promotion. Once the selection board concludes, the official board report and the promotion recommended list are sent to the Service Secretary. In reviewing, the Secretary's staff compares the promotion recommended list to the promotion decisions projected by a separate machine learning algorithm that was not provided to the promotion board members. Differences are noted and substantial differences are checked for irregular patterns.*

No single system is infallible. The objective here would not be to build an algorithm that outperforms a promotion board, with all the nuances of promotion decisions that take careful deliberation and judgement. Its intent would be as an additional check in the spirit of having multiple independent models and processes that serve as checks on each other. An entirely separate model can be used as an auditing and validation tool, where the auditing can cross-check another model's predictions. In this case, the first model is an information input into the board's decision-making process, and a second model serves as a check on the decision-making process itself. That said, although a separate model can add a degree of robustness—particularly in highlighting divergent predictions—both models may be wrong in the same way, preventing problems from being detected because the predictions do not diverge.

One challenge here is that promotion decisions are based in part on demand requirements, which can change from year to year. Consequently, promotion standards can ebb and flow somewhat. An algorithm may capture those who are clearly promotable and those who clearly are not; however, it will likely struggle with those who rank in the middle. An algorithm that uses data

on the performance of previous cohorts would likewise need to account for nuanced changes to the promotion standards.

2. Traceable

The DOD's AI Ethical Principles call for all stages of the machine learning pipeline to be traceable. The principle states that, "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."

When incorporating machine learning into existing policies, traceability lies not only in the machine itself, but also in its use. For some processes that are currently opaque to the service member, machine learning may be a tool to provide more transparency than existing processes allow. This transparency may not necessarily be a product of the machine learning, per se, but rather the emphasis on having well-documented, data-driven processes that can inform decision making. This aspect of traceability is distinct from that of the machine learning itself.

To facilitate traceability of the actual machine learning model, developers should provide relevant documentation of processes for designing the model and interpreting its outputs. There should be a fully documented, reproducible pipeline, stretching from the raw data, to any data preprocessing steps, to the modeling, and then finally to the results. Documentation should also be available on the types of robustness checks that have been done on the model. Reviewers and stakeholders should also be able to suggest alternative robustness checks. Depending on the specific use case, robustness checks may need to focus more on minimizing false positives or false negatives. Assumptions about any thresholds used for transforming continuous output probabilities into distinct categorical classifications should be explicit.

a. Officer Promotions

Incorporating machine learning into the promotions process can occur while maintaining similar levels of transparency and auditability as the current process allows. Currently, selection board processes are clearly outlined in policy. Existing practices place strong limits on the visibility of the promotion board's proceedings. Except perhaps under highly extraordinary circumstances, officers under consideration for promotion have no visibility into these proceedings. Deliberations are maintained only in a confidential board record, which is maintained for a certain period of time for auditing purposes. Officers not selected for promotion may appeal decisions but not view the report on the board's decision-making process. When considering how to introduce machine learning into this process, leaders would need to consider whether this would also limit an officer's right to request information about any algorithms used to summarize information about them.

Currently, selection boards consider a body of information clearly specified in policy, which officers under consideration for promotion have reviewed prior to the board convening. Current DOD policy allows for automated computer summaries of information. However, machine learning models may enable the incorporation of a larger body of data than has been used to date. If the DOD considers the use of additional data acceptable, it will also need to consider whether any practices need to be changed to ensure that officers are reviewing all relevant data within their file. Because current policy allows officers to see all promotion materials that the board considers, do they have a right to receive a copy of their own summary metrics?

For context, there are some current performance metrics with limited visibility for service members. One argument for being opaque is that service members whose performance is below average, but still acceptable enough to be valuable to the military, may exit service prematurely if they knew how they ranked relative to their peers. A counterargument to being opaque is that if service members have little visibility into how the military sees them, and if they do not receive feedback on areas in which they need to develop, then service members will not have sufficient direction to improve. Ideally, visibility would be coupled with meaningful and actionable guidance on how to improve.

b. Better Forecasting and Programming for Training Slots

In the vignette of forecasting and budgeting for training slots, the Air Force was considering replacing a legacy process for predicting training slots with a machine learning approach. When incorporating machine learning into a process, at a minimum, the new process should be at least as traceable as the status quo in following the end-to-end process of ingesting data into a model and producing an output. In some cases, machine learning may provide further traceability than was previously available (for instance, if analyses had been done in ad-hoc spreadsheets with minimal documentation that only the creator of the spreadsheet could navigate).

Machine learning processes may also introduce traceability in the sense of standardization. If many different offices are creating forecasts for their own particular use, each with their own methodology, then there may be limited traceability at the aggregate level in knowing how each forecast was derived. A central, standardized method for producing forecasts may be more traceable than decentralized forecasts. Yet, performance must also be considered. The central method should perform at least as well as the decentralized forecasts.

Stakeholders likewise need to consider the audience for the model's traceability. The DOD principle of transparency refers to "relevant personnel." In the training slots vignette, leaders faced some pushback from stakeholders who expressed skepticism about the model. Such skepticism may be based in part on not receiving detailed information about how the model arrives at its predictions. Traceability is essential to a model's auditability, but it may also be important to achieving buy-in and appropriate use.

3. Reliable

Monitoring the safety and effectiveness of machine learning is critical to maintaining legal, moral, and ethical standards. The DOD stipulates that "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." Reliability in this sense spans a model's life-cycle. The data and model must reside in a secure environment with adequate safeguards against adversarial attacks — a particular concern for personnel data and models that have both privacy and readiness sensitivities. The model should be built and used for specific use cases. The model also needs ongoing support and testing. Defining a use case and testing that the model meets an objective implies that the model must reliably satisfy a given standard of performance.

a. Recruiting and Selection for Special Forces

The MARSOC example highlights the importance of comparing the reliability of a machine learning-based approach to traditional approaches for recruiting and selecting Special Forces candidates. Currently, MARSOC applies rigorous standards in all phases of the Special Forces selection process. To meet the reliability standard of being an effective tool, it is critical that any application of machine learning enable MARSOC to maintain those same high standards. MARSOC should assess whether individuals the model identifies truly have the needed skills and abilities. Ongoing tests may be needed to determine this. Machine learning tools can be rolled out in such a way that allows for monitoring and evaluation of their effectiveness, such as by comparing predictions to actual outcomes before employing the predictions.

If many individuals that the model predicted do not go on to complete the traditional MARSOC selection process, decision-makers should work closely with developers to determine whether the model requires refinement or if there are other reasons for this discrepancy. Any algorithm will inevitably produce false positives and false negatives, but the goal for reliability should be in meeting a given standard of performance. Stakeholders will need to develop metrics for evaluating model performance and determine performance thresholds above which the model needs to perform in order to be effectively used.

b. Officer Promotions

Monitoring the application of the machine learning tool can also help ensure it is used as intended. In the fictional vignette on officer promotions, metrics summarizing an officer's performance and potential are incorporated into an existing promotion review board process. Some of these metrics rely on machine learning. Reliability of machine learning in this context is related to how decision-makers use the model's predictions within this process. Prior to deploying the model, the intended use should be clear to all stakeholders. Leaders should update policies and provide guidance to address how the metrics (including those based on machine learning predictions) should and should not be incorporated into the process.

In the officer promotions vignette, the intended use of the machine learning model is to synthesize data and provide selection board members an additional piece of information to inform their deliberations; the machine learning model is not the decision-maker. Many factors contribute to promotion selections, not all of which will be reflected in a model's predictions, and selection board members should still use their own judgement about promotions.

The salience of the metrics in the promotion materials will likely impact how much board members rely on them to inform their opinions of the candidates. In implementing the metrics, it may be worth evaluating different options for providing metrics to board members, with the goal of providing useful information, but not having the metrics be the sole factor considered in the evaluations. For instance, instead of providing the summary metrics upfront along with the other information on the candidates, the summary metrics could be reserved until later in the review process so that the board members have a chance to review the other materials first. The metrics might also be reserved for providing additional information for just a subset of candidates, such as borderline candidates.

To support board members in appropriately using the metrics, board members must be able to interpret the information correctly and understand any potential limitations of the underlying models. Developers should provide training materials so that the original purpose and limitations of the underlying models can be easily communicated to users. Promotion board members should have enough information about how the metrics are calculated so that they can understand what the metrics are measuring, what data are being considered in calculating the metrics, and, if desired, information about the algorithm and its performance metrics. The promotion board members should also be briefed on known or potential shortcomings of the metrics. Such training needs to be short and easy to understand for those with limited technical backgrounds.

c. Targeted Retention Interventions

Retention decisions are likely influenced by myriad factors, such as personal preferences, health considerations, availability of professional opportunities within and outside of the military, and national sentiment toward military service. As internal and external influences on these decisions shift over time, the original model may become less reliable in targeting individuals who have a high likelihood or option value of exiting service. Implementation of the model may also reveal that predictions are less reliable over different time horizons or for different career fields or populations. This necessitates testing and refinement throughout the model's life-cycle. The frequency with which the model needs to be refreshed will vary depending on a number of factors. Developers should be able to conduct performance checks for specific populations of interest on an ongoing basis to account for this.

Actions that the military takes in response to the model can also shift how those targeted by the model behave. If retention incentives are offered solely to those who are forecast to leave, those who planned to stay in service may shift their behavior to appear as if they would leave in order to receive an incentive. Hence, the reliability of the model is impacted by this feedback loop of

actions taken in response to the model. Implementation plans would therefore need to account for that. Incorporating the predictions into a broader decision rubric may help to mitigate negative aspects of a feedback loop. The fictional vignette, for instance, has a rubric that combines the individual retention forecasts with information on officer performance, as well as forecast shortages and surpluses for given skills.

4. Governable

Governability refers to the ability to control the use of the machine learning model, and particularly the ability to disengage the model when its use is no longer desired. DOD's AI Ethical Principles state that, "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." In applying this principle to the use of machine learning in personnel policy, it is important to understand what those unintended consequences may be. Most directly, a model that produces incorrect predictions will have the unintended consequence of inefficient or potentially harmful human resource management. Beyond this, incorporating machine learning into existing processes can also affect factors such as trust, cohesion, fairness, and service member wellbeing.

a. Targeted Retention Interventions

Any attempt to influence service members' decision-making regarding whether to remain in or leave the military risks unintended consequences, regardless of whether the intervention involves machine learning. For example, service members may have expectations about the types of personal information that are (and are not) available to the military and the ways in which the military can use that information. A lack of trust may develop if service members perceive that their privacy has been violated. Even if the capabilities of the machine learning model use appropriate data and are employed in appropriate ways, a lack of trust — if pervasive enough — could more than undo any potential benefits of the algorithm.

A lack of trust, or feelings of underappreciation could also develop if individuals not selected for the targeted intervention feel that they were passed over unfairly. Leaders need to be sensitive to public perception and be able to sensitively address concerns. This may include providing additional educational material on the model and how it is being used, but it may also include altering the use of a model or entirely disengaging its use for certain applications.

It is also possible that after deploying the model that the quality of its forecasts substantially decreases for individuals in a particular demographic group or career field — placing those individuals at a disadvantage for receiving a targeted intervention. Quality checks are needed on an ongoing basis to assess the reliability of the model. Additionally, there needs to be a governance mechanism for determining when interventions based on the model may need to be adapted due to the model's failure to meet given quality checks. This requires some balance. Models are fallible, but human decision making in the absence of good data and analytics can also be flawed.

Performance expectations for a model should be appropriately benchmarked against existing processes and any shortcomings that the status quo may embody.

b. Recruiting and Selection for Special Forces

There are a range of unintended consequences that could result from changing the Special Forces selection process. There are reliability concerns about whether machine learning can predict training completion. The model may do well at predicting that individuals with certain characteristics are likely to succeed, but it may do poorly at predicting for non-standard candidates that may also succeed. Predictions focused on a particular type of candidate may have a homogenizing effect. This may have unintended consequences for limiting the skill and experience profile to a standard type, excluding a breadth of talents and backgrounds that may be necessary ingredients to successful special forces units.

There are also questions of what, if anything, is lost by recruits not completing the full Assessment and Selection process. Even if machine learning can improve the efficiency of this process, there may be benefits to having a less efficient process. Perhaps, for example, the current training program develops character traits and capabilities necessary to uphold core values and serve effectively in special forces units. It may be that these traits can still be developed to some degree through an abbreviated training.

In this fictional vignette, MARSOC should monitor whether an abbreviated selection process has any unintended effects after rollout of the model and weigh costs and benefits. If the model's costs outweigh its benefits, MARSOC would need to consider whether it is appropriate to disengage the model or to revise its use. Given that MARSOC selection is a resource-intensive process, it may not be feasible to simply switch back to the old process, if machine learning fails. As stakeholders plan for the deployment of models, this consideration may inform the extent to which they decide machine learning can supplant certain components of the selection process that may be more difficult to restore.

5. Responsible

Although "responsible" is the first of the five DOD AI Ethical Principles, we address it last since it encapsulates the other four principles. It is described as follows: "DOD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities." Given this and the other definitions, there are few topics that uniquely fall under the heading of "responsible" that do not already fall under the purview of the other four principles. In that sense, "responsible" may be thought of (in a positive light) as a parent principle that drives the other principles. However, if the goal is to have a parsimonious set of principles, "responsible" could be viewed (in a negative light) as an unfocused catch-all. It is overly broad to be clearly actionable. That said, one topic that may be relevant here is data privacy and the responsible use of sensitive data.

a. Responsible Use of Sensitive Data

Personnel data can be inherently sensitive due to privacy concerns. For members of the military, personnel data have an added layer of sensitivity since they collectively describe elements of defense readiness. Depending on the precise nature of personnel data, they may be subject to the protections of Personally Identifiable Information (PII), Protected Health Information (PHI), or various classified protections.

Working with such data requires a safe and secure information system (as noted under the principle of reliability). The sensitivity of data elements can vary considerably: ranging from the benign to the highly sensitive. Some data may only become sensitive as it is combined with other information. For instance, knowing a particular stateside duty location or promotion date may not be sensitive by itself, but knowing extensive details about the careers of one or more individuals might become sensitive.

The question of what data elements may be legally or ethically permissible to use as inputs to a model does not have a definitive answer. There are multiple factors at play:

- Information technology risk: If the data or model were compromised through an adversarial attack or other mishap, are there certain data elements (or combinations of elements) that are particularly sensitive?
- Potential for violating privacy norms: Can the intended use of the model be perceived as violating a social norm of privacy?
- Potential for perceived or actual bias: Do the data inputs suggest a preferential treatment of one or more demographic groups at the expense of another?

In the vignette on improving forecasting and programming for training slots, person-level data are used as inputs. However, the final output is an aggregate number of training slots needed within a training pipeline. Because the final output is far removed from the individual, even if more sensitive data on individuals are used within the algorithm, there is little concern of an undue privacy violation or suggestion of bias. The upper bound on the sensitivity of the data to include in this case is consequently set primarily by the information technology risk and the potential for a data leak.

Bias and privacy norms play a more prominent role in the other vignettes. The MARSOC vignette, for instance, involves machine learning assisted recruiting efforts to target individuals in the Marine Corps who may be a good fit for the Special Forces mission. A concern of potential bias may arise if a distinct demographic of Marines who would likely qualify for Special Forces was omitted from the recruiting efforts. It would be wise to assess the outputs of the machine learning predictions along various demographic dimensions to check whether any demographic groups are represented at lower than expected rates. Out of further caution, it may also be appropriate to exclude information from the model that may be perceived as potentially biased —

whether or not the data elements do in fact introduce bias. However, as discussed in Chapter 1, this is often not straightforward because some data elements can proxy for others.

The vignette on targeted retention interventions noted the uncanny timing of an officer receiving an attractive intervention right at the time that she was considering schooling and a career outside the military. Isolated coincidences are one thing, but if individuals feel that the military can consistently predict too many things that they have not disclosed, that may lead to a sense of distrust in the amount of surveillance that the military has over their lives. Privacy norms may likewise be violated if machine learning is used to predict information that is not directly related to their professional career.

For instance, although having a child may impact a service member's retention probability, *explicitly* predicting when a female service member is likely to become pregnant as an intermediate step to predicting her retention probability is arguably a privacy violation. The harder question is whether a machine learning algorithm should be able to *implicitly* deduce pregnancy patterns within an algorithm that predicted retention probabilities for female service members.

When deciding whether to include or exclude data elements — due to potential privacy, bias, or security issues — it is important to remember that the choice need not be a binary decision. A third option is to mask more sensitive data elements to some degree. This can be done, for instance, by binning elements into fewer identifiable groups (e.g., using 3-digit U.S. Postal Service ZIP codes instead of 5-digit ones), or by deliberately adding noise to the data (e.g., perturbing values so that the true value is only preserved a certain percentage of the time).

7. Additional Recommendations

The ethical risks of machine learning are real. But so are the potential ethical benefits of machine learning applications. Ethical codes and regulations should take these risks seriously, but do so in a way that balances development costs, as well as the ethical consequences of *not* developing methods that may synthesize information better or more transparently.[103]

Tools to identify these risks and mitigate them are in flux. Approaches that overly fixate on particular identification and mitigation methods may be counterproductive. Instead, we recommend a general ethical framework and general processes that should be particularized on a case-by-case basis. We conclude with a few recommendations for mitigating legal, moral, and ethical risks of machine learning in the context of personnel management.

A. Agile Development

Machine learning models cannot be developed, deployed, and simply forgotten. The predictive power of machine learning models trained on historical data can diminish over time. The underlying relationships in the historic data on which the predictions are based can shift in response to policy changes, evolving cultural norms, geopolitical events, economic conditions, and a variety of other factors.

Given that operational models and the data underlying them need to be updated periodically, the full pipeline for producing models should be version-controlled. This includes storing metadata on the precise data elements used to produce a model, any code used to process the data and run the model, and any user-specified parameters for the model. Ideally, the ways in which the model output are incorporated into any business management or other decision-making processes should also be documented. This enables transparency into both the underlying modeling and the application of the modeling result.

Tools that enable interpretation of the model by outside reviewers should be prepared in advance to facilitate prompt investigation of legal, moral, or ethical concerns. Metrics tracking model performance must be maintained and evaluated on an ongoing basis.

A suite of tests should be developed to check for potential ethical risks (what may be termed *ethical unit tests*), and these tests should be run on each version of the model prior to

[103] Such a summation of benefits and costs is an inherently consequentialist perspective.

deployment.[104] Ideally, the development of these tests should begin early in the life-cycle of the machine learning application.[105] Although it is not possible to run comprehensive statistical tests for ethical violations, it may be possible to flag some warning signs where additional attention is warranted. For example, a disparate prediction (or disparate impact) may not, by itself, be a legal or ethical violation. However, combined with other circumstances, a large disparate prediction could indicate a potential legal or ethical risk. Disparate predictions over relevant groups should be reviewed.[106]

Legal, moral, and ethical review processes must operate within a dynamic environment. To avoid lengthy delays in responding to issues and concerns as they arise, processes should be designed to be proactive and flexible.

B. Legal, Moral, and Ethical Review and Oversight

In adopting and using machine learning models in personnel management contexts, it may be appropriate to establish ad hoc or standing review bodies to provide oversight for how machine learning models are used. Such reviews may be especially appropriate when incorporating machine learning into processes that are already subject to heightened levels of scrutiny, such as promotion boards. For applications that are less sensitive to the career progression of individual service members, such as budget projections or aggregate staffing forecasts, it may be adequate to simply subject the model to robust peer-review or some form of a verification, validation, and accreditation (VV&A) process. In any case, it is important to maintain clear documentation of the input data, the model, and how the output data is used in a management process.[107]

When convening review bodies, it is important to include members with expertise in machine learning, law, and the relevant personnel management process. The body should also avoid conflicts of interest by excluding or limiting the membership of those who have a significant vested interest in the use of the model. The review body must maintain enough independence so that it

---

[104] The U.S. Census Bureau's efforts to ensure non-disclosure of identifiable data may provide a useful framework for thinking about how a suite of ethical unit tests might be developed and implemented.

[105] The software development concept of DevSecOps (Development, Security, and Operations) cultivates a security-focused culture throughout the entire development life-cycle, as opposed to just have security checks at the final stage of development. An analogous concept, which might be termed DevEthOps (Development, Ethics, and Operations), is needed to incorporate ethical considerations throughout the development life-cycle.

[106] It is extremely unlikely that predictions from a personnel model will be identical across groups unless the model were engineered to generate identical predictions (such as through disparate treatment). In this sense a disparate prediction is inevitable in the absence of disparate treatment. It may be appropriate to measure things such as the size of the disparate impact (particularly in comparison to prior policy) and to reflect on any likely underlying problems that can and should be remediated.

[107] Machine learning techniques are simply algorithmic tools. As such, they are only as sensitive as the issue to which they are applied. Every application of machine learning within the personnel management context need not be subject to legal, moral, and ethical review (just as every regression model, simulation, or other algorithm need not automatically be subject to legal, moral, and ethical review).

can provide a genuine review. Code, data, and a viable execution environment should be accessible to the review body so that it can perform independent tests of the model. To facilitate peer review and deeper investigation, the model should preferably be released to an open-source community. The review body should facilitate timely feedback and accommodate agile development.

Poorly implemented review processes misallocate accountability. One risk is that the reviewing body will be held accountable for legal or ethical harms but will not be held accountable for forgone benefits. If such a body can unilaterally block applications of machine learning models, it may be destructively conservative. By symmetry, those commissioning the machine learning application may feel that a formal approval absolves them of responsibility for legal or ethical harms. However, the complex and dynamic risks in personnel policy require that they remain engaged. The balance between accountability and immunity must include considerations for both the upsides and downsides.

To weigh risks appropriately, reviewers must be provided with appropriate resources and processes. When resources are limited, reviews should focus on applications that are likely to have the greatest risk.

C. Rolling out New Processes

The ethical risks of using machine learning models in specific personnel management applications may not be immediately apparent. It may therefore be prudent to deploy the new process incrementally—also an agile development principle—to provide learning opportunities while also reducing the scope for potential harm. Early stages of the rollout should follow experimental design principles where appropriate to enable the impact of the tool on specific processes to be assessed.[108] Evaluation data should be collected to evaluate performance and ethical concerns. Discussions with subordinates and those affected by the new process should solicit perceptions of fairness, equity, and any other recommendations.

Proactively allaying concerns can be vital to facilitate stakeholder buy-in. This is especially true when incorporating machine learning models into potentially sensitive processes. As appropriate, documentation can be developed to inform senior DOD leaders, Congress, service members, and the public. This can include a description of the former process, how machine learning may impact the new process, what elements are considered in the machine learning model and how, and what safeguards are in place to ensure that underlying goals (such as equality of opportunity) are maintained. Open communication can help to mitigate adverse consequences.

Appropriate mechanisms should be in place to redress harms. This includes responding to public perceptions. At least initially, personnel may be suspicious of machine learning

[108] In cases where the deployment is designed as a formal research study that can contribute to generalizable knowledge (as opposed to, for instance, an internal assessment of organizational processes), and may therefore be subject to an Institutional Review Board (IRB), the IRB may help to fill the need for legal, moral, and ethical oversight.

applications. Promptly and publicly responding to concerns will help to maintain trust. This also includes redressing harms to individuals. Depending on context, this may involve an appeal process for those adversely affected by decisions informed by machine learning models. Although the applicability of antidiscrimination law to machine learning in personnel policy is uncertain, proactive use of administrative processes for redressing grievances may help to bolster trust and dissuade adverse reactions.

In addition to legal, moral, and ethical concerns, a thoughtful deployment will also consider other potential unexpected outcomes. For instance, in the fictional vignette on officer promotions, a summary metric provided to the board may be based on an algorithm designed to examine early career characteristics of successful field grade officers. Once the scores are provided to and used by a board, it may be that the board chooses those with particular types of combat arms experience at a much higher rate than previously. If this results in an undesirable balance of skill sets and occupational specialties, other actions may be required to restore the desired balance.

D. Document Uses and Limitations

The documentation on a model should include information on its intended use and any known limitations. This documentation should also include a discussion of possible legal and ethical risks for the processes that the model is intended to support, as well as any strategies for mitigating risks. Documentation may need to be developed both at a technical level and at a level accessible to users. Lay users should be able to understand what the model does and does not consider.

Suppose, for example, that a machine learning model helps a recruiter to target recruits by predicting which recruits have the highest probability of successfully completing Initial Entry Training. Documentation, such as the following, could help the recruiter to balance the model's predictions against information that is not included in the model.

- *"The prediction you have been given indicates that the recruit is unlikely to complete Initial Entry Training. One possible concern is that the recruit has a high body mass index (BMI), which is often in indicator of obesity. However, it can also indicate that the person is very muscular. Does the recruit appear to be physically fit?"*
- *"The prediction you have been given does not consider the recruit's moral fiber. You should use your judgement, along with testimonials from coaches, teachers, and others, to weigh the recruit's potential for service against the predicted probability of attrition risk."*

A central risk of machine learning is that its application will prioritize what can be easily measured over what cannot. For instance, manpower and recruiting requirements can be carefully tracked and measured. Other goals are less easily quantified. Training requirements and readiness metrics reflect anticipated needs, but may fail to anticipate lapses in mobilization timetables or performance in theater. Enlistee and officer quality affect readiness, but quality can be difficult to measure.

In light of these and other broad, holistic considerations, decision makers must be equipped to appropriately consider machine learning predictions—balancing the information that the model is using against additional information that the decision maker may have that is not included in the algorithm. What information does the model *not* account for that it ideally would? What information does the model account for that it ideally would not?

Decision makers should feel empowered to account for contextual information that the machine learning predictions did not have, so long as that additional information is itself legal, moral, and ethical to consider. Incorporating information from machine learning into operational processes should be done carefully to mitigate unintended consequences.

This page is intentionally blank.

Appendix A. Illustrations

Tables

Figures

This page is intentionally blank.

Appendix B. References

Bauer, William A. "Virtuous vs. utilitarian artificial moral agents." AI & SOCIETY 35, no. 1 (2020): 263–271.

Bersin, Josh, "Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age," *Forbes*, February 17, 2013, https://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/?sh=42cb4b2a4cd0.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. "Building classifiers with independency constraints." In 2009 IEEE International Conference on Data Mining Workshops, pp. 13–18. IEEE, 2009.

Calmon, Flavio P., Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "Optimized pre-processing for discrimination prevention." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3995–4004. 2017.

Celis, Elisa, Lingxiao Huang, Vijay Keswani, Nisheeth K. Vishnoi, "Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees." arXiv:1806.06055v3.

Daly, Sarah, Metin Toksoz-Exley, "Expanding the Ethical AI Conversation: Virtue and its Implications for the Development and Use of Artificial Intelligence Enabled Capabilities [Shortened]," Institute for Defense Analyses, 2022.

Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

Elish, Madeleine Clare, "Who is Responsible when Autonomous Systems Fail?" *Center for International Governance Innovation*, June 15, 2020, https://www.cigionline.org/articles/who-responsible-when-autonomous-systems-fail.

Elish, Madeleine Clare, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interactions," *Engaging Science, Technology, and Society* 5 (2019): 40–60.

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar, "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," Report submitted to the Administrative Conference of the United States,

February 2020, p. 76, https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

Gutmann, Amy, and Dennis F. Thompson. "What Deliberative Democracy Means," in Why deliberative democracy? Princeton University Press, 2009.

Hagendorff, Thilo. "The ethics of AI ethics: An evaluation of guidelines." Minds and Machines (2020): 1–22.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016): 3315–3323.

Hibbard, Bill. "Ethical artificial intelligence." arXiv preprint arXiv:1411.1373 (2014).

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1, no. 9 (2019): 389–399.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016).

Lipton, Zachary C., Alexandra Chouldechova, and Julian McAuley. "Does mitigating ML's impact disparity require treatment disparity?" In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8136–8146, 2018.

Lockwood, Julie, Rachel Augustine, Joseph King. *Identifying Correlates of Navy Line Officer Retention and Promotion among various Demographic Groups Machine Learning for Hypothesis Generation WEAI 2021*. IDA Paper NS P-22655. Alexandria, VA: Institute for Defense Analyses, June 2021.

Lockwood, Julie, Alan Gelder, Matthew Goldberg, Jennifer Brooks, George Prugh. *Leveraging Machine Learning in Defense Analyses*. IDA Paper P-13174. Alexandria, VA: Institute for Defense Analyses, May 2020.

Lockwood, Julie, Joseph King, Rachel Augustine. *Explaining Differences in Predicted O-5 Promotion Outcomes by Race and Gender among Naval Officers*. IDA Paper P-20452. Alexandria, VA: Institute for Defense Analyses, December 2020.

Mallon, David, Jeff Moir, Robert Straub, "People Analytics: Gaining Speed," Deloitte, 2016, https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2016/people-analytics-in-hr-analytics-teams.html.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." arXiv preprint arXiv:1908.09635 (2019).

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

Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." Nature Machine Intelligence (2019): 501–507.

Nestler, Scott. "Data Ethics and Decision-Making." Presentation at the Institute for Defense Analyses, July 14, 2020.

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

Spain, Everett S. "Reinventing the Leader-Selection Process: The U.S. Army's new approach to managing talent." *Harvard Business Review*. November–December 2020. https://hbr.org/2020/11/reinventing-the-leader-selection-process.

Verma, Sahil, and Julia Rubin. "Fairness definitions explained." In *2018 IEEE/ACM International workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

This page is intentionally blank.

Appendix C. Abbreviations

AI	Artificial Intelligence
BMI	Body Mass Index
DOD	Department of Defense
DP	Definitely Promote
FY	Fiscal Year
IDA	Institute for Defense Analyses
IEEE	Institute of Electrical and Electronics Engineers
MARSOC	Marine Forces Special Operations Command
MBA	Master of Business Administration
ML	Machine Learning
NDAA	National Defense Authorization Act
PHI	Protected Health Information
PII	Personally Identifiable Information
ROC	Receiver Operating Characteristics
ROTC	Reserve Officer Training Corp
SEPTA	Southeastern Pennsylvania Transit Authority
SHAP	SHapley Additive exPlanations
VV&A	Verification, Validation, and Accreditation

This page is intentionally blank.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
XX-04-2023	Final	March 2020 - May 2022

4. TITLE AND SUBTITLE

Legal, Moral, and Ethical Implications of Machine Learning for Personnel Management

5a. CONTRACT NO.

HQ0034-14-D-0001

5b. GRANT NO.

5c. PROGRAM ELEMENT NO(S).

6. AUTHOR(S)

Alan Gelder
Julie Lockwood
Cullen Roberts
Ashlie Williams
Kathleen Conley
Rachel Augustine

5d. PROJECT NO.

5e. TASK NO.

BE-6-4311

5f. WORK UNIT NO.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NO.

IDA Paper P-33087
Log: H 22-000189

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD(P&R)
1400 Defense Pentagon,
Arlington, VA 22202

10. SPONSOR'S / MONITOR'S ACRONYM(S)

OUSD(P&R)

11. SPONSOR'S / MONITOR'S REPORT NO(S).

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Machine learning is increasingly used to synthesize and harness data in support of decision-making processes. While machine learning models can be more powerful than other analytic techniques, they also have the potential to aggravate the risk that information is misused in decision making. This Institute for Defense Analyses paper attempts to clarify the foreseeable legal, moral, and ethical risks of machine learning — as well as what can be done to mitigate these risks — when applied to personnel management processes. Although the primary focus is on the military setting, the underlying lessons apply broadly to personnel management in a variety of contexts. Building on key legal principles, normative and applied ethics, and technological capabilities and constraints, we provide practical considerations and recommendations for using machine learning to support personnel management processes.

15. SUBJECT TERMS

Machine learning, Artificial Intelligence, Legal, Moral, Ethical, Personnel Management, Data Science Ethics, Military Personnel

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NO. OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Lernes Hebert
U	U	U	U	108	19b. TELEPHONE NUMBER (Include Area Code) (703) 571-0114

This page is intentionally blank.