Research Report

DAVID SCHULKER, MATTHEW WALSH, AVERY CALKINS, MONIQUE GRAHAM, CHERYL K. MONTEMAYOR, ALBERT A. ROBBERT, SEAN ROBSON, CLAUDE MESSAN SETODJI, JOSHUA SNOKE, JOSHUA WILLIAMS, LI ANG ZHANG

Leveraging Machine Learning to Improve Human Resource Management

Volume 1, Key Findings and Recommendations for Policymakers

KEY FINDINGS

- To generate business value by meaningfully contributing to human resource management (HRM) process efficiency and workforce capabilities, the U.S. Deparrtment of the Air Force (DAF) must first grow a machine learning (ML) project portfolio made up of technically feasible projects that address near-term and future HRM needs.
- To effectively develop ML systems, the DAF must first specify HRM objectives and then select modes of decision support that meet those objectives.
- To act legally, ethically, and responsibly, the DAF must test candidate systems to ensure that they are *safe*—that is, accurate, fair, and explainable.
- To overcome inertia, the DAF must pursue transition pathways that involve gradually increasing the degree of ML influence or, alternatively, gradually increasing the significance of the HRM processes at stake.

The national security environment poses strategic challenges for human resource management (HRM) policies and systems. Then-Air Force Chief of Staff Gen Charles Brown (Chairman of the Joint Chiefs of Staff as of this writing) captured this challenge succinctly in the first of his "action orders" in September 2020:

> Past success is no guarantee of future performance. The [U.S. Air Force] must ensure the future force reflects the identity and attributes required for success in the high-end fight. Tomorrow's Airmen must be organized, trained, and equipped to succeed in the

Recommendations

To continue to evaluate and leverage data technologies in HRM, the U.S. Department of the Air Force (DAF) must continuously undertake the four sets of activities listed below. The recommendations corresponding to the four sets of activities and summarized here are developed in the remainder of this report.

- **Manage innovation constantly:** Require a well-formulated business case and technical feasibility assessment before projects can move forward. Adopt a portfolio approach to managing complexity.
- **Develop effectively:** Begin the design process with priority objectives and consider multiple modes of decision assistance. Prioritize development of machine learning (ML) systems that automatically summarize narrative records as a mode of decision support.
- **Implement safely:** Use the accurate-fair-explainable framework described in this report to create tailored designs that safely meet objectives. Publish acceptable limits for safety criteria in different classes of use cases to encourage adoption.
- **Transition strategically:** Regulate the stakes of the HRM decision and the amount of influence allotted to the ML system to find an implementation that balances value and risk. Apply ML systems to limited cases before gradually expanding their scope and consequence.

> most challenging and lethal combat scenarios since World War II. (Brown, 2022)

Commensurate with General Brown's action order, an independent committee organized by the National Academies of Sciences, Engineering, and Medicine, which included experts in the fields of industrial and organizational psychology, economics, and information science, also concluded that the DAF must adapt its HRM system to meet the challenges of the national security environment. In its final report, the committee chose to emphasize *how* to strengthen the HRM system through a "flight plan" made up of implementable actions, grouped under three major priorities: the data priority, the airmen priority, and the research priority (National Academies of Sciences, Engineering, and Medicine, 2020).

Within the data priority, the final report places strong emphasis on the use of new or emerging technologies (Table 1 lists and defines some of the most common such technologies for reference). The report lists "digital qualitative methods, including technical capabilities in natural language processing, text mining, and other emerging advances" as a "key enabler/driver" of the data priority. The recommended actions for the data priority include a call for "expand[ing] the capability for predictive and prescriptive analytics to support personnel decision processes" as well as "identify[ing] opportunities to incorporate artificial intelligence and other technologies to improve data flow, granularity of decisions, and speed of interactions." Finally, the actions highlight the particular potential of natural language processing (NLP) to help extract information from the text of HRM records, stating that the DAF should "consider a constellation of technologies, including machine learning and natural language processing, to better comprehend collected data and inform decision-making" (National Academies of Sciences, Engineering, and Medicine, 2020). This emphasis on adopting new data technologies further nests within a broader policy context of U.S. Department of Defense (DoD) and DAF strategies and directives for suborganizations to better leverage data in decision-making (e.g., see DoD, 2018).

Furthermore, these recommendations are part of a broader context of technology development in the HRM and broader workplace domains. For example, IBM has published research testifying to the business value of its own artificial intelligence applications in HRM, which include "candidate attraction, hiring, learning, compensation, career management, and HR support" (Guenole and Feinzig, 2018). There also continues to be a large and dynamic market for technology-driven HRM service offerings with embedded artificial intelligence (Josh Bersin Company, 2021; Budhwar et al., 2022).

TABLE 1
Common Data Technologies and Definitions

Term	Definition	U.S. Air Force Example
Supervised ML	Techniques that seek to transform data inputs into a predicted output, based on example pairs of inputs and outputs	Using data on past trainees to create a model that predicts training attrition risk based on trainee characteristics, such as fitness and aptitude
Unsupervised ML	Techniques that seek to transform and organize data inputs without regard to producing a particular output, usually with the goal of understanding or simplifying data inputs	Using an algorithm to identify groups of similar positions for tracking and workforce planning
Reinforcement learning (RL)	Techniques that seek to learn predictive rules using a reward function representing feedback from the system's environment (rather than based on example pairs of inputs and outputs)	Developing algorithms capable of piloting an F-16 in simulated combat for the Defense Advanced Research Projects Agency's AlphaDogfight Trials (Pope et al., 2021)
NLP	Cross-cutting term for ML techniques that generate application functionality from models of human language	Using sentiment analysis to automatically classify open-ended survey responses as positive or negative
Optimization	Finding a set of inputs that produces the best possible result from a model	Finding a set of selective reenlistment bonus multipliers that achieves retention objectives at the lowest possible cost

Together, these strategies and findings lead to two conclusions: (1) To successfully meet strategic objectives, the DAF must improve the way it develops and manages human capital, and (2) the DAF should explore ways to leverage data management and analysis technologies, such as ML and NLP, to realize these capability improvements. These conclusions connect directly to the research objective for this study, which is to develop ML-based decision-support methods and tools that help HRM panel members or managers process and understand textual performance records.

Given that the DAF (and DoD, more generally) is making broad investments in improving its data and analysis technology, it is worthwhile to study how to leverage those investments for the greatest impact. However, we do not intend this objective to imply that these technologies are the only solution to HRM challenges. From the perspective of any particular project, functional managers should assess the totality of possible improvements when considering the potential value of introducing technological solutions.

Abbreviations

DAF	U.S. Department of the Air Force
DEDB	developmental education designation board
DoD	U.S. Department of Defense
HR	human resources
HRM	human resource management
ML	machine learning
NLP	natural language processing
OPR	officer performance report
PReSS	Performance Records Scoring System

Our Research Approach Spans the Life Cycle of Technology Adoption

Research on broader adoption of artificial intelligence (much of which is likely powered by ML methods) shows that organizations, including those with deep technical expertise, face unique challenges in the HRM domain. Survey results from private industry show that HRM is among the business functions in which adoption is the lowest (Chui et al., 2021). To shed light on the potential reasons for relatively low adoption, Tambe, Cappelli, and Yakubovich (2019) conducted a workshop with workforce ana-

lytics chiefs from 20 major U.S. corporations. Their findings reveal barriers that relate to the inherent complexity of measuring HRM outcomes coupled with data constraints, ethical or legal unknowns that are difficult to work through, and possible negative impacts on affected employees (Tambe, Cappelli, and Yakubovich, 2019). Professional guidelines for HRM also emphasize that organizations might want to avoid selection procedures that are vulnerable to legal challenges, or those that employees or other stakeholders view as controversial (Society for Industrial Organizational Psychology, 2018). These risks loom especially large given that labor law considers procedures that have an adverse impact on protected groups to be discriminatory by default unless organizations can justify the procedures through rigorous validation (Code of Federal Regulations, Title 29, Part 1607).

While there have been many successful demonstrations that apply such technologies as ML and NLP to DAF HRM problem sets (Robson et al., 2022; Schulker, Harrington, et al., 2021; Schulker, Lim, et al., 2021; Walsh et al., 2021), the research on adoption shows that identifying, developing, and implementing valuable decision-support systems is more than an analytic challenge. Therefore, in pursuing our research objective of developing decision-support methods and tools for DAF HRM processes, we organized our research tasks around four main aspects of the technology adoption life cycle:

1. How can the DAF build and oversee a portfolio of research and development projects exploring the use of ML in HRM?
2. How can the DAF effectively develop decision-support systems based on NLP?
3. How can the DAF test decision-support systems to confirm they are safe to use in decisionmaking?
4. How can the DAF strategically transition systems for operational use once they are developed and tested?

Our theory of change is that leveraging data technologies for strategic impact requires DAF decisionmakers to systematically select the right mix of projects, effectively execute the development of selected projects, establish procedures for testing decision-support systems to address ethical and legal unknowns, and successfully transition systems into use in a way that is acceptable to the adopting organizations.

To guide decisionmaking in the functions that support this life cycle, we developed a series of tailored reports to discuss different elements of the theory of change (see Table 2). Each report provides details on research methods and data supporting the topic, as well as results and detailed findings. This

TABLE 2
Outline of Report Series

Volume Number	Report Title	Report Purpose
1	*Leveraging Machine Learning to Improve Human Resource Management:* Vol. 1, *Key Findings and Recommendations for Policymakers* (Schulker, Walsh, et al., 2024)	Overview for senior leaders
2	*Machine Learning in Air Force Human Resource Management:* Vol. 2, *A Framework for Vetting Use Cases with Example Applications* (Walsh et al., 2023)	Framework for how to prioritize ML projects
3	*The Personnel Records Scoring System:* Vol. 3, *A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking* (Schulker, Williams, et al., 2024)	Technical report on scoring officer records
4	*Safe Use of Machine Learning for Air Force Human Resource Management:* Vol. 4, *Evaluation Framework and Use Cases* (Snoke et al., 2024)	Case study approach to ensure safety of ML systems
5	*Machine Learning–Enabled Recommendations for the Air Force Officer Assignment System:* Vol. 5 (Calkins et al., 2024)	ML system to inform officer assignments

NOTE: Current report is highlighted.

report is intended to extract the strategic findings and recommendations from across the tasks, organized around the theory of change, to inform senior leaders in their ongoing decisions as they follow strategic guidance and pursue applications of ML in the HRM domain. The other reports contain additional findings and recommendations that are not included in this strategic summary.

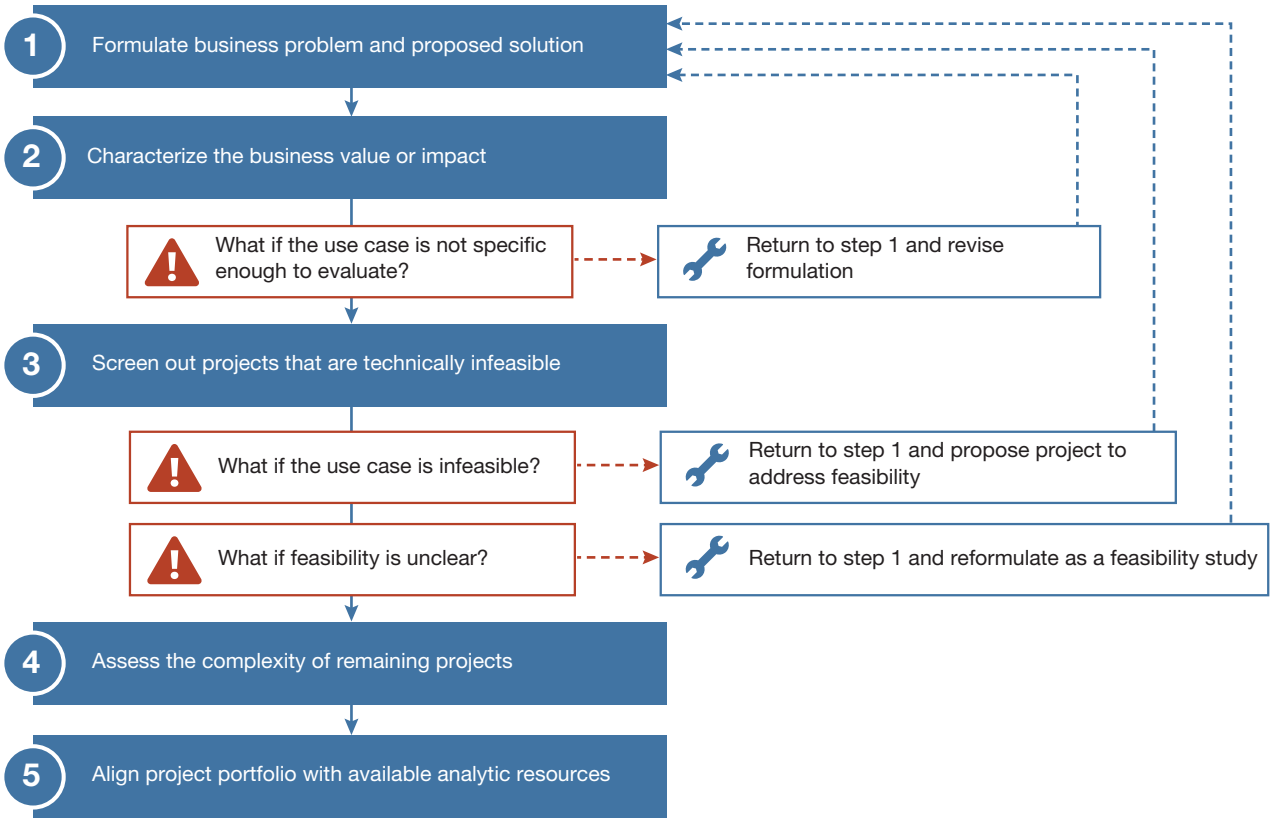
Managing the Innovation Portfolio

The past decade has seen significant breakthroughs in ML and NLP. For example, deep neural networks can accurately forecast HRM outcomes, reinforcement learning models can efficiently control dynamic systems, and large language models can fluently mimic human dialogue (Vaswani et al., 2017). The enthusiasm around these breakthroughs is reflected in the business literature and in the sizable government and commercial investments in ML.

Organizations are no doubt expecting to garner a large return on ML investments. Yet the business literature also reveals that many organizations fail to realize significant business value from such investments. These organizations add technical staff, build infrastructure, and run demonstration projects—all things that the DAF does. But their efforts are ad hoc (Fountaine, McCarthy, and Saleh, 2019). The result is that (1) ML projects do not address the most significant business opportunities, or (2) they do not become mature enough to transition from demonstrations to full implementation.

To avoid this outcome, DAF decisionmakers must become experts at selecting the right ML projects. Figure 1 displays a core set of evidence-based steps to guide portfolio management. This framework is not intended to replace processes that DAF analytic decisionmakers already use for evaluating and prioritizing potential projects. Rather, elements of this framework can be incorporated into these processes, and the framework can provide senior

FIGURE 1
Framework for Selecting a Portfolio of ML Projects for HRM



leaders with a consistent picture of a wide variety of potential HRM ML applications. In Volume 2 (Walsh et al., 2024), we further suggest an innovation dashboard that tracks projects along with the outcomes of the steps in Figure 1, so that decisionmakers are aware of ideas that have already been proposed, along with their potential sources of business value, technical feasibility, and both technical and nontechnical (e.g., privacy, cultural barriers to adoption) aspects of complexity.

In reviewing the effectiveness of DAF processes for choosing ML projects for HRM, decisionmakers should reflect on the following recommendations, based on our assessment of the most-distinctive elements of the selection framework.

Recommendation 1a. Require a well-formulated business case and technical feasibility assessment before projects can move forward

The first three steps in the framework seek to address common reasons that technical projects fail to produce business value. Enthusiasm over recent breakthroughs in ML makes it tempting to begin projects without a clear understanding of business value, with the hope that the agile development process will reveal a path to success (Bryar and Carr, 2021). Decisionmakers should resist this temptation by sending projects that lack a fully formulated business case, including a precise plan for how the project will deliver business value, to the back of the queue. This practice does not prohibit exploratory projects because decisionmakers can go beyond direct business value and consider the value of discovery and the value of creating new future decision options (Shore, 2022).

Requiring a plausible description of the business value of a technical project at the outset enables more-informed decisions and, in the long run, better outcomes. Furthermore, forcing organizations to articulate the business value of a project can be a useful tool for holding them accountable for undertaking the change management steps necessary to follow through with implementation, such as training their employees to use the tools and eliminating obsolete manual processes. To assist with implementing this recommendation, Volume 2 (Walsh et al., 2024) defines the four main areas of HRM business value and suggests key metrics for each. These areas include value created through process improvement, enhancing workforce skills and performance, increasing workforce motivation, and improving opportunities to use existing skills and motivation.

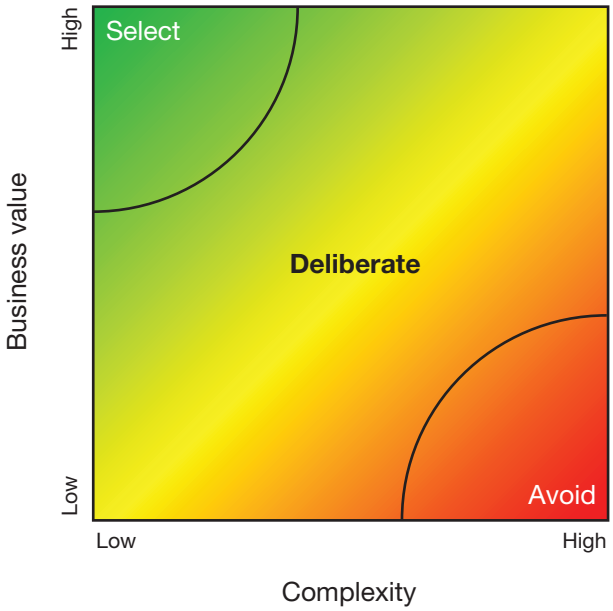
Aside from requiring a fully formulated business case, decisionmakers should request assurances about potential projects' technical feasibility. This step does not ensure project success, but it may screen out projects that, due to technical infeasibility, are destined to fail. Notably, in our assessment of nearly 20 wide-ranging ML projects, suitable algorithmic methods existed for most. The primary limitation pertained to the availability of adequate data inputs and outcome measures, both of which echo back to the private-sector research on barriers to ML adoption in HRM.

Recommendation 1b. Adopt a portfolio approach to managing complexity

A consistent theme in the research literature is the need for organizations to manage the complexity of projects and, implicitly, the risk of failure. Projects can be technically complex, depending on the data, modeling, or software challenges inherent in creating a system that meets HRM objectives. However, other sources of complexity (i.e., nontechnical) include the policy structures in place that govern how the system might be used, stakeholders with competing interests, and other workforce considerations, such as cultural adaptation and workforce retraining requirements that the new system drives.

Rather than simply prioritizing projects in order of predicted business value or level of complexity, decisionmakers should manage these attributes with a portfolio approach (Figure 2). Once high-value, low-complexity activities (top left of Figure 3) are underway and low-value, high-complexity activities (bottom right) are weeded out, the remaining resources should fund a mix of proposals along the complexity-value spectrum, including high-risk, high-reward approaches. The conventional wisdom from private-sector companies is that about 70 per-

FIGURE 2
Complexity-Value Matrix for Decisionmaking



cent of an innovation portfolio should support core business functions and the remaining 30 percent should support more-transformative initiatives (Nagji and Tuff, 2012). This value-versus-complexity scoring technique allows decisionmakers to ensure that most resources deliver incremental wins and to begin shaping the culture of adoption, while some high-risk, high-reward efforts open the DAF up to transformative changes.

Executing Development Projects for Decision-Support Tools

Once a project has been added to the portfolio, decisionmakers must develop the ML system by either using organic analytic resources or supervising development (if working with partner organizations). Those involved in the design of such systems will immediately run into the challenge of prioritizing among a wide range of possible features and design parameters. Should the system automate a decision, provide inputs to human decisionmakers, or interact with the decision process in some other way? What inputs do the human decisionmakers need, and how effective are candidate ML systems at delivering these inputs? What are the risks of different types of decision support, given the current level of functionality available for different candidate systems? Success in designing decision-support systems that contribute to HRM value demands an ordered process of thinking through these questions.

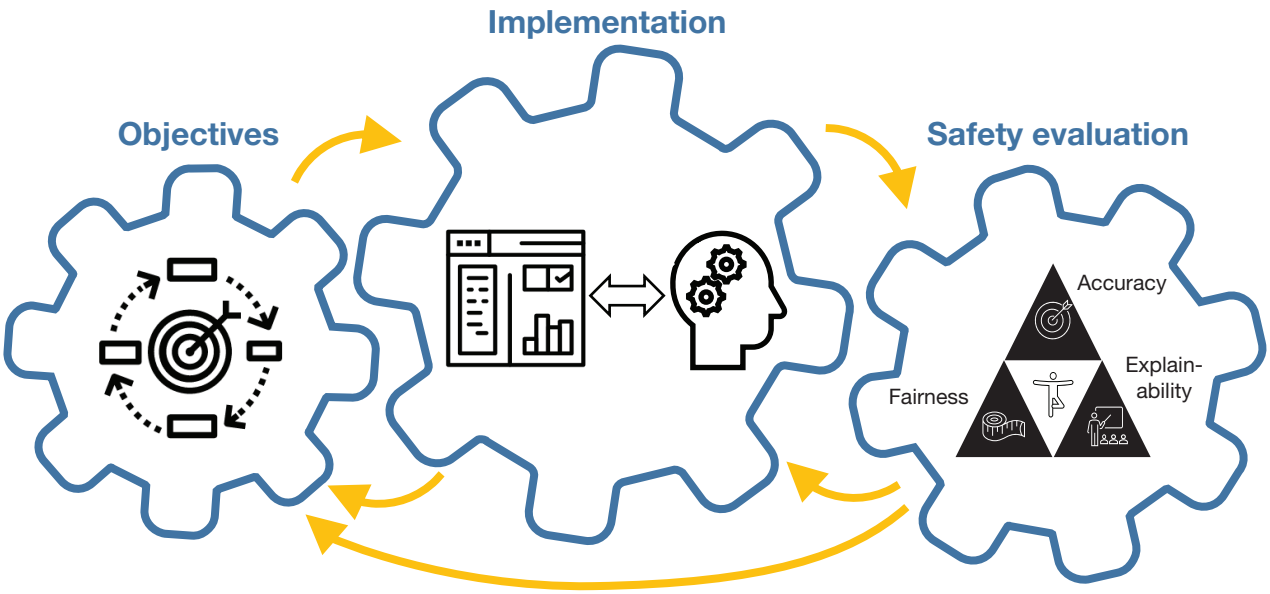
Figure 3 presents a framework for conceptualizing and designing ML systems for HRM. The first insight behind the framework is that the system design is inseparable from the highest-priority objectives of the system. The HRM objectives help designers select from among the many possible ways that ML can support an HRM decision process. The implementation design, in turn, affects how the system is evaluated. For example, a system that automates a decision can be evaluated based on its accuracy (among other important criteria), whereas a system that provides inputs to human decisionmakers must be judged based on the accuracy of the inputs and how they affect the overall decision outcome. If the system fails to satisfy safety criteria, the implementation design must be modified until designers can arrive at a system that contributes value to the HRM objective and can meet safety parameters. Our recommended safety criteria of accuracy, fairness, and explainability are discussed further in the following section of this report and in Volume 4 of this series (Snoke et al., 2024), which focuses on safe and equitable use of ML for HRM.

To explore ways to effectively develop ML and NLP-based systems for HRM decision support, we undertook a project intended to serve as an archetype for systems that support human resources (HR) managers who review records as an input and render various decisions (e.g., for developmental education or promotion purposes). We designed and tested a performance scoring system, which we dubbed the Performance Records Scoring System (PReSS). As described in subsequent sections of this report, we followed the framework outlined in Figure 3 to design, implement, and evaluate PReSS.

The engine of PReSS is a set of ML models that take an individual officer's record of officer performance reports (OPRs), break it into key terms and phrases, and then use the presence or absence of those terms and phrases to predict either a board

FIGURE 3

Framework for Selecting and Evaluating ML System Implementation Design



NOTE: This framework is conceptually similar to standard processes for solving data science problems. For an example, see Chapman et al., 2000.

score, in the case of developmental education designation boards (DEDBs), or a promotion probability to the grades of O-5 and O-6. The ML models learn the value of individual terms and phrases by correlating them with past DEDB scores and promotion outcomes. A completed model can then apply those values to the terms and phrases in a new officer record to arrive at a prediction. This analytic process is described in more detail in Schulker, Lim, et al. (2021), and in Volume 3 (Schulker et al., 2024) of this series.

The primary output of the PReSS prototype is a general performance summary of the officer's record of OPRs that quantifies the performance level in the record and extracts key phrases based on their impact on the estimated score. Such a summary could support a variety of HR decision processes that require performance as an input. Figure 4 shows a sample of these outputs for a fictional OPR containing both performance-related text and unrelated text from Air Force Instruction 1-1. PReSS takes a set of OPRs and produces the following:

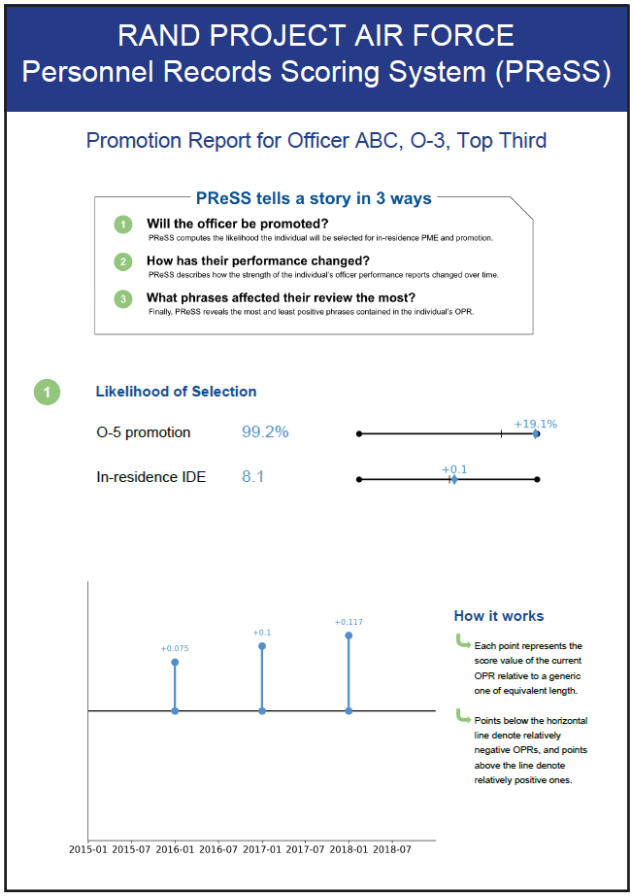
1. a quantitative prediction—the overall summary at the top left of Figure 4—of how upcoming DEDBs and promotion boards would view the record
2. a time-series plot of the quality of each OPR—performance over time at the bottom left of Figure 4—revealing the underlying trend in performance
3. a text summary of the most-positive and -negative statements contained in the record, shaded by the degree of positivity or negativity, which is the section of Figure 4 on the bottom right.

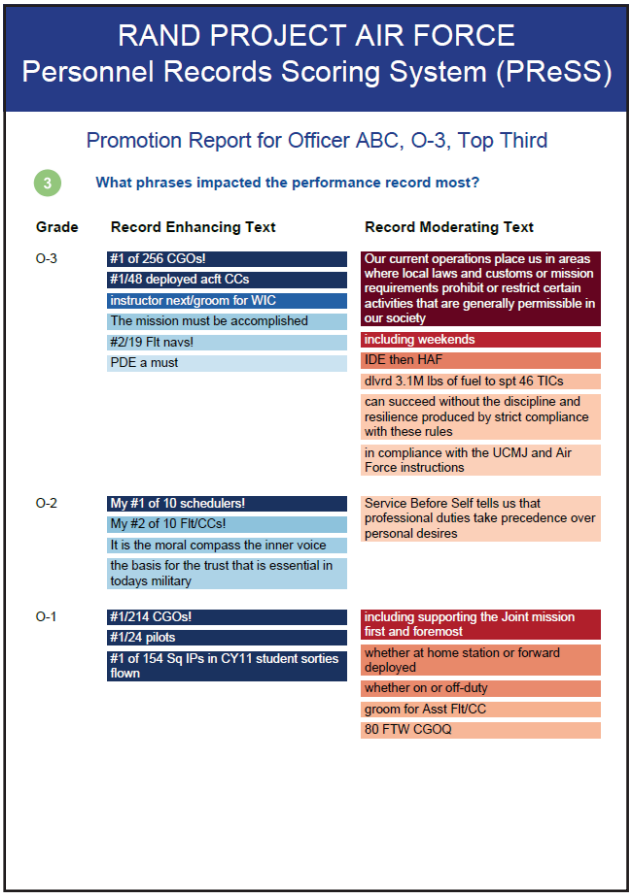
Volume 3 (Schulker et al., 2024) describes the development process and provides a detailed overview of the PReSS general performance summary and the methods behind it.

In developing PReSS, we uncovered principles of interest to policymakers who oversee development projects for HRM decision-support systems, leading to the following recommendations related to designing these systems.

FIGURE 4

Sample Outputs from PReSS





NOTE: The PReSS methodology compares overall records, individual OPRs, and phrases within OPRs with a vector of equivalent length that represents the average term use throughout the data. Thus, the report identifies the positive or negative impact of an OPR or a phrase by comparing the predicted score including the phrase in the record with the predicted score if the record contained the same number of "average-looking" terms. In practice, very few records contain statements describing negative performance. Instead, raters differentiate performance by replacing highly recognizable signals with vaguely positive placeholders. Thus, lukewarm statements such as "groom for Asst Flt/CC" could indicate average or below-average performance. The sample report contains the following abbreviations (in the order listed): Intermediate Developmental Education (IDE), OPR, Company Grade Officer (CGO), Commander (CC), Flight (Flt), Weapons Instructor Course (WIC), Professional Developmental Education (PDE), Headquarters Air Force (HAF), Troops in Contact (TIC), Uniform Code of Military Justice (UCMJ), Instructor Pilot (IP), Calendar Year (CY), Assistant (Asst), Fighter Wing (FTW), and Company Grade Officer of the Quarter (CGOQ).

Recommendation 2a. Begin the design process with priority objectives and consider multiple modes of decision assistance

Early in a development project, it becomes clear that there are many design options for incorporating ML-based inputs into decisionmaking. Designs vary in terms of timing (e.g., before versus after a human has formulated a decision) and degree of influence (e.g., recommending an option versus directing attention to important features). Throughout this series, we focus on five main *design implementations* to integrate ML decision-support systems with board processes:

1. *Decide*. The ML system scores HRM records and automatically arrives at decisions without human decisionmaker involvement.
2. *Recommend*. The ML system provides recommendations to human decisionmakers as additional inputs.
3. *Score*. The ML system provides scores to humans as additional inputs.

4. *Summarize*. The ML system automatically summarizes OPRs for human decisionmakers.
5. *Audit*. The ML system flags irregular cases for human decisionmakers to review as part of an auditing process.

Rather than beginning with a particular implementation in mind, such as automatically scoring records (i.e., option 3 above), we recommend that the design process start with identifying the priority objectives of the ML system (refer to Figure 3). Different combinations of objectives call for different design implementations, as summarized in Table 3. The objectives also point to potential measures of effectiveness to evaluate the process. For example, if the goal is to reduce workload, the system should reduce the number of human decisionmakers or the amount of time they spend scoring records. Alternatively, if the goal is to improve human decisionmaking, the system should contribute to increased quality of decisions, measured by evidence that the decisions better contribute to important HRM outcomes.

We further find that it is beneficial to consider multiple modes of decision support to increase available options for meeting HRM objectives safely, as discussed in the next section of this report. To give one example, decisionmakers can use PReSS to completely automate the DEDB selection process, which would maximize the objective of reducing human workload by eliminating human boards entirely. But we show in Volume 4 of this series of reports (Snoke et al., 2024) that the current DEDB models are likely not accurate enough to meet safety considerations under an automatic design. Instead, an implementation in which human decisionmakers use PreSS general performance summaries or recommendations as inputs can still reduce workload but are more likely to satisfy safety criteria by keeping a human in control of the process.

Recommendation 2b. Prioritize development of ML systems that automatically summarize narrative records as a mode of decision support

The majority of a person's HR record is free-form text—lists of assignments, descriptions of duties, syllabi of educational courses, and summaries of key accomplishments. HR records also contain prequantified, interpretable personnel attributes that are useful for management, such as years of experience, order of merit, or promotion test scores. While the latter type of information is easier to process and use in a model or visualization, the former is also needed to make fully informed HRM decisions.

Historically, the DAF handles decisions that require deliberate review of textual inputs with a manual review or scoring process performed by experienced officers, and many HRM decisions fall into this category. This means that the general devel-

TABLE 3
Alignment Between System Implementation Designs and ML Objectives

Objectives	Decide	Recommend	Score	Summarize	Audit
Provide feedback	–	–	–	++	–
Increase transparency	+	–	–	++	+
Standardize processes	++	+	++	+	+
Improve human decisionmaking	–	++	+	++	+
Reduce workload	++	++	–	+	–
Advance DAF priorities	+	++	++	+	+

++ = high alignment; + = moderate alignment; – = low alignment.

opment techniques behind PReSS could apply in many other areas.

Of the various design implementations that we considered for supporting manual reviews, the "summarize" implementation is the most general. This was the only design that is moderately or highly aligned with all HRM objectives that we considered. Automated summaries are highly useful for providing feedback, increasing transparency, and improving the accuracy of human decisions, and they are at least moderately useful for standardization, reducing human workload, and steering decisions toward DAF priorities. At the same time, the summarize implementation maintains a high level of human control over the decision process, so it is more likely than other designs to meet safety criteria. In fact, a summary highlights elements of the text that the system considers important, and therefore, it is a type of explanation for the system's decisions. Thus, summaries can be a useful companion tool for helping managers understand the model's outputs in any of the other design implementations that we considered.

For all these reasons, the DAF should continue to invest in capabilities to automatically extract and summarize information from free-form text, as PReSS does. Note that such summaries would need to be evaluated using safety criteria, discussed in the following section and in Volume 4 of the series (Snoke et al., 2024).

Demonstrating the Safety of ML Systems

HRM decisions affect humans and the health of the future force. Thus, the DAF must adopt a "first, do no harm" principle for making major changes to decision processes. To implement PReSS-like applications to support HRM decision processes, decisionmakers must be able to confidently and credibly assert that the implementation is safe. This requires concrete definitions of the elements of safety and metrics for assessing them.

With increased investment in ML has come a large body of research and policy papers aiming to provide normative guidance for using ML (and artificial intelligence more broadly) responsibly and

The DAF should continue to invest in capabilities to automatically extract and summarize information from free-form text, as PReSS does.

ethically. These policies include DoD's own Responsible Artificial Intelligence Strategy ethical principles, which state that use of artificial intelligence needs to be responsible, equitable, traceable, reliable, and governable. These and other ethical principles[1] are essential to overseeing and regulating the development and implementation of decision-support systems. Current rules and frameworks for protecting member privacy, for instance, would continue to apply to any development project.

Three principles are especially relevant to testing a system during development and deployment, and in Volume 4 (Snoke et al., 2024), we demonstrate how to apply them to our prototype decision-support system. Our testing framework posits that safety considerations require ML systems to be accurate, fair, and explainable. We colloquially refer to these criteria as the *iron triangle*. There are likely minimum thresholds for each criterion that will depend on the application. Further, the need to balance the three, as opposed to focusing solely on one, is represented by the balancing person pictured within the evaluation circle of Figure 3.

- *Accuracy* means that the ML system or the model that it contains correctly predicts the outcome of interest with a high probability[2]
- *Fairness* means that the ML system treats subgroups equivalently

Each element of the framework presents potential risks that must be weighed given the objective and intended use of the system.

- *Explainability* means that a human can understand the factors and relationships that led to the ML system's outcome.

These safety criteria are sometimes in tension with one another. To increase fairness, designers might place constraints on the system that reduce its accuracy or explainability. For example, career and performance outcomes have historically differed across demographic groups (Asch, Miller, and Malchiodi, 2012). To increase fairness, system designers might blind ML models to protected attributes like race and gender, or they might use less explainable technical constraints to force the system to produce similar predictions across groups. To increase explainability, system designers may use more interpretable—but less flexible—modeling approaches, which could affect both accuracy and fairness. Testing necessarily involves balancing accuracy, fairness, and explainability to arrive at a design that meets HRM objectives and legal and ethical constraints.

Each element of the framework presents potential risks that must be weighed given the objective and intended use of the system. A failure of accuracy could occur if the model draws on inaccurate or inappropriate data,[3] or if the model makes incorrect predictions (either when predicting decisions, such as whether to select an individual, or predicting quantitative values, such as the quality score of a record). Any such failure risks harming individuals if the system contributes to errors in decisionmaking.

Regarding fairness, it is important to note that there is no single definition of *fair*, and it is often not possible to satisfy competing types of fairness. Thus, institutions must choose a definition to move forward with testing. In our application, we differentiate between procedural fairness, which ensures that an HRM process or algorithm treats members of different subgroups the same, and outcome fairness, which examines the model or process outcomes for bias.[4] Under the umbrella of outcome fairness, potential metrics examine whether model predictions differ by group, possibly conditional on other factors, and whether model errors skew systematically against particular groups.[5] The consequences of failing to meet fairness criteria are severe and include the ethical risk of discriminating against protected categories of employees, as well as the legal risk of violating regulations prohibiting discrimination.

Finally, explainability is critical for achieving HRM objectives, because humans might ignore or misuse systems if they do not understand how they contribute to better decisionmaking. Furthermore, defining *explainability* is inseparable from the intended audience, because different types of users will require different levels of explanations. Designers can consider using models that are inherently interpretable to increase explainability, and they can also conduct human-in-the-loop testing to gauge how well people understand the functionality of the system.

Volume 4 describes the process for defining and measuring accuracy, fairness, and explainability of ML systems for HRM (Snoke et al., 2024). Drawing on our findings from using this framework to evaluate PReSS for the use case of officer selection boards, we offer the following policy recommendations for evaluating ML-based decision-support systems more generally.

Recommendation 3a. Use the accurate-fair-explainable framework to create tailored designs that safely meet objectives

There are not universal definitions for *accuracy*, *fairness*, and *explainability*, and satisfying one definition does not ensure that the ML system satisfies others. For example, fairness could mean that the ML system yields equal outcomes for different demographic

groups or that it yields equal outcomes for otherwise identical individuals from different demographic groups. Thus, to permit concrete evaluations of ML systems, decisionmakers must first select from agreed-upon definitions of *accuracy*, *fairness*, and *explainability*, and they must specify the relative importance of each given a particular application.

Consider one subset of our selection board use cases as a brief illustration. Existing selection board processes begin by having panels score all records. Then, records that are close to the selection threshold (a.k.a., *in the gray*) receive another review before the board renders a final decision. PReSS could replace the first round of scoring, freeing up panel members to focus exclusively on the gray-zone records, thus saving labor while affording panels additional time to focus on the hardest cases (Figure 5).

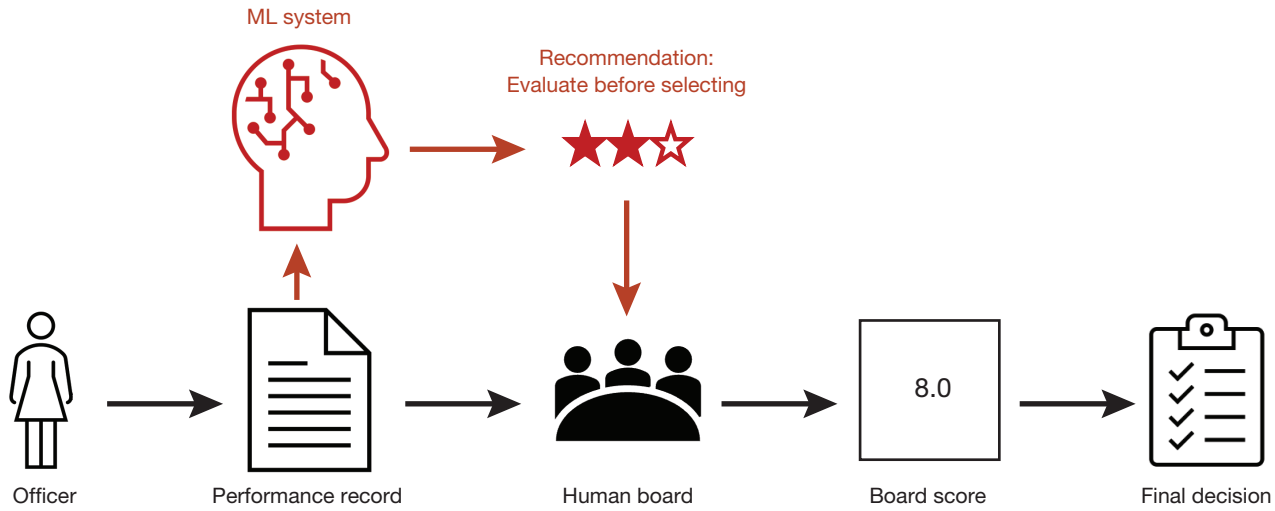
To arrive at a safe design for this system, decisionmakers must first define *accuracy*, *fairness*, and *explainability*. *Accuracy* means that the system would agree with a human-centric board process a high percentage of the time. We consider multiple definitions of *fairness* in Volume 4 (Snoke et al., 2024) but, for simplicity, assume that we want to meet the strictest definition of fairness: *independence*, which requires that the system outcomes are the same for members of different subgroups—that is, final selection rates for PReSS should be the same for all demographic groups. Finally, *explainability* means that human board members can understand the model inputs and outputs and the procedure for generating predictions.

Our results when we analyze promotion boards show promise. For both O-5 and O-6 selections, PReSS models have high accuracy, they are fair along the gender dimension, and they are inherently interpretable.[6]

However, if we were unable to achieve a safe version of the *decide* implementation for the first stage of this process, this would not be the end of the design process. We could select an alternate design implementation that maintains human control of that stage of the process, or we could even use the system in an auditing capacity, in which it would not directly influence the decision process but could still contribute to some of the objectives. This limited summary of how to apply the accurate-fair-explainable criteria illustrates how the framework equips decisionmakers and developers to articulate safety criteria and examine systems through a safety lens.

FIGURE 5

Example ML System Implementation Design in Which System Recommendation Automates Part of the Board Decision



NOTE: The ML system scores officers' OPRs and divides records into three tiers: select, evaluate before selecting, and do not select. Human boards score records labeled *evaluate before selecting,* and they accept system recommendations for records labeled *select* and *do not select.*

Recommendation 3b. Publish acceptable limits for safety criteria in different classes of use cases as a means of encouraging adoption

One factor limiting greater adoption of ML in the HRM domain is uncertainty over the level of safety required. How safe is safe enough? Here, policies might encourage adoption by publishing acceptable limits or benchmarks for safety criteria in different classes of use cases. For instance, Title 29 of the Code of Federal Regulations states that, in the civilian labor context, cases in which a subgroup selection rate is at or below 80 percent of the rate for the most-selected group are generally considered evidence of adverse impact (Code of Federal Regulations, Title 29, Part 1607). While the law states that some smaller differences could still constitute adverse impact, the benchmark provides needed clarity to employers when designing recruitment and selection processes. If the DAF were to adopt such rules for different classes of use cases, it would empower HRM practitioners to adopt ML systems, knowing that they have met agreed-upon safety standards.

Finding the Right Transition Pathway for an ML System

The ability to meet safety criteria is a no-latitude requirement for any use case. This reality might seem to be in tension with the principles of agile software development, which stress releasing increments as early as possible to reap the maximum value over time and to generate feedback crucial for rapid improvements (Shore, 2022). In the case of ML for HRM, decisionmakers can resolve this tension by further tailoring their system implementation according to a decision-support system's level of maturity. Early releases of a system can target lower-stakes use cases with less-strict safety requirements, or they can exert less influence over outcomes by retaining humans in the decision process. As the system improves over time and garners higher trust and confidence from users, decisionmakers can alter the system's implementation scope and increase its contribution to HRM objectives. The following recommendations will help decisionmakers find transition pathways for new ML systems.

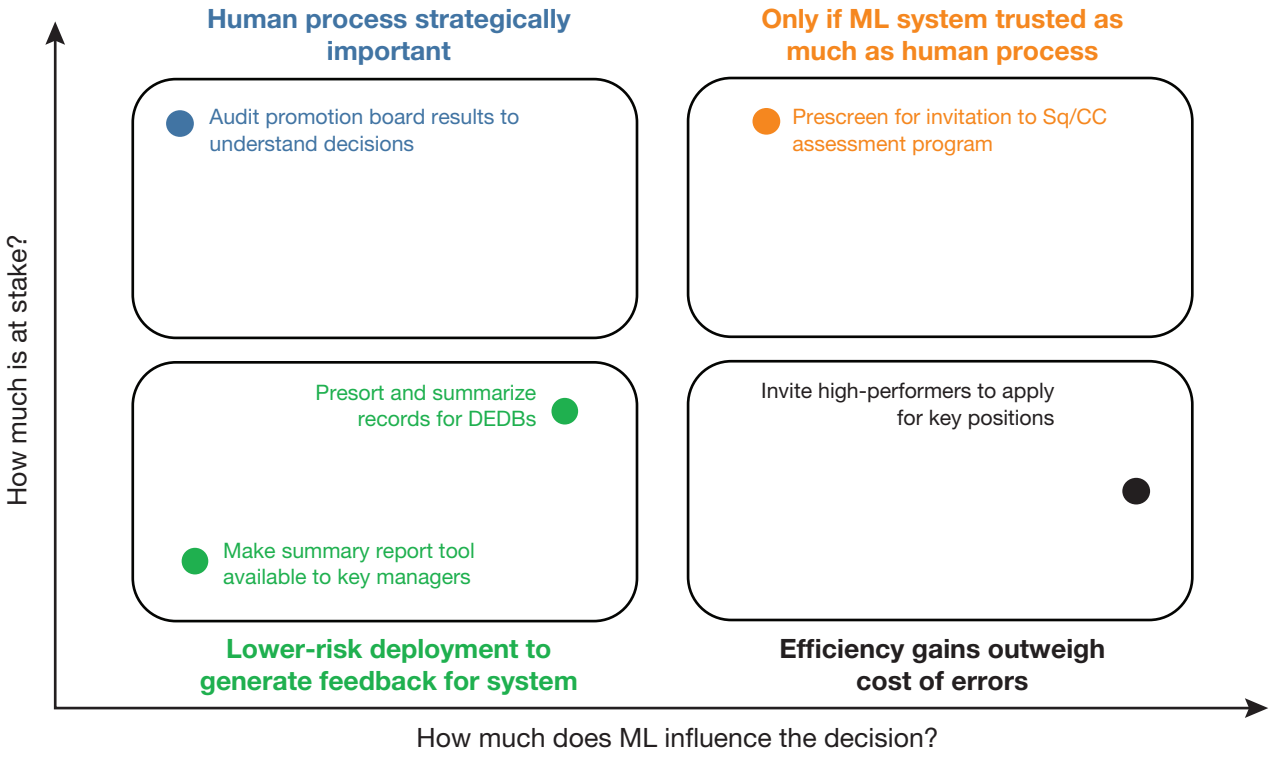
Recommendation 4a. Regulate the stakes of the HRM decision and the amount of influence allotted to the ML system to find an implementation that balances business value and risk

We suggest two primary dimensions that determine the best transition pathway for an ML system in the HRM domain: the stakes of the decision and the amount of influence the ML system has on the decision. Different combinations of these dimensions lead to the four zones depicted in Figure 6. Implementations in which both the stakes of the decision and the influence of the ML system are low represent low-risk deployment cases for minimally viable products (lower left quadrant). These implementations are valuable because they offer opportunities to expose users to an imperfect system so that it can be improved. Decisionmakers may then extend low-influence ML systems to high-stakes processes (upper-left quadrant). This is appropriate when human ownership of the process is of strategic importance, such as for promotion boards.

Alternatively, decisionmakers may increase the degree of ML influence for low-stakes processes (lower-right quadrant). This is appropriate when efficiency gains from greater reliance on the ML system outweigh the costs of errors. Ultimately, decisionmakers may adopt implementations in which both the stakes of the decision and the influence of the ML system are high (upper-right quadrant). This is appropriate only if decisionmakers trust the ML system at least as much as they trust existing human-centric processes. The dots within each quadrant further illustrate the framework using variations on how PReSS can be employed.

FIGURE 6

Framework for Selecting and Evaluating ML System Implementation Design, with Example PReSS Implementations



NOTE: Sq/CC = squadron commander.

Recommendation 4b. Apply ML systems to limited cases before gradually expanding their scope and consequence

HRM decisionmakers often do not face a go/no-go decision with respect to whether to use an ML system. If a particular implementation is deemed too risky, they can select from potentially less capable but safer designs for the initial deployment and gradually increase either the ML influence or the outcome stakes to generate greater business value as the system proves itself trustworthy. For example, using PReSS models to fully automate a selection decision is likely too risky at this stage in its development. But decisionmakers can reduce the risk by adopting the less influential "recommend" or even the "audit" design implementations to begin to generate business value while continuing to improve the system. Alternatively, they could identify use cases for PReSS that could have a greater influence—but on lower-stakes decisions. They could even consider a minimally influential implementation for low-stakes decisions, such as providing the PReSS general performance summary capability to managers or individuals during a testing phase.

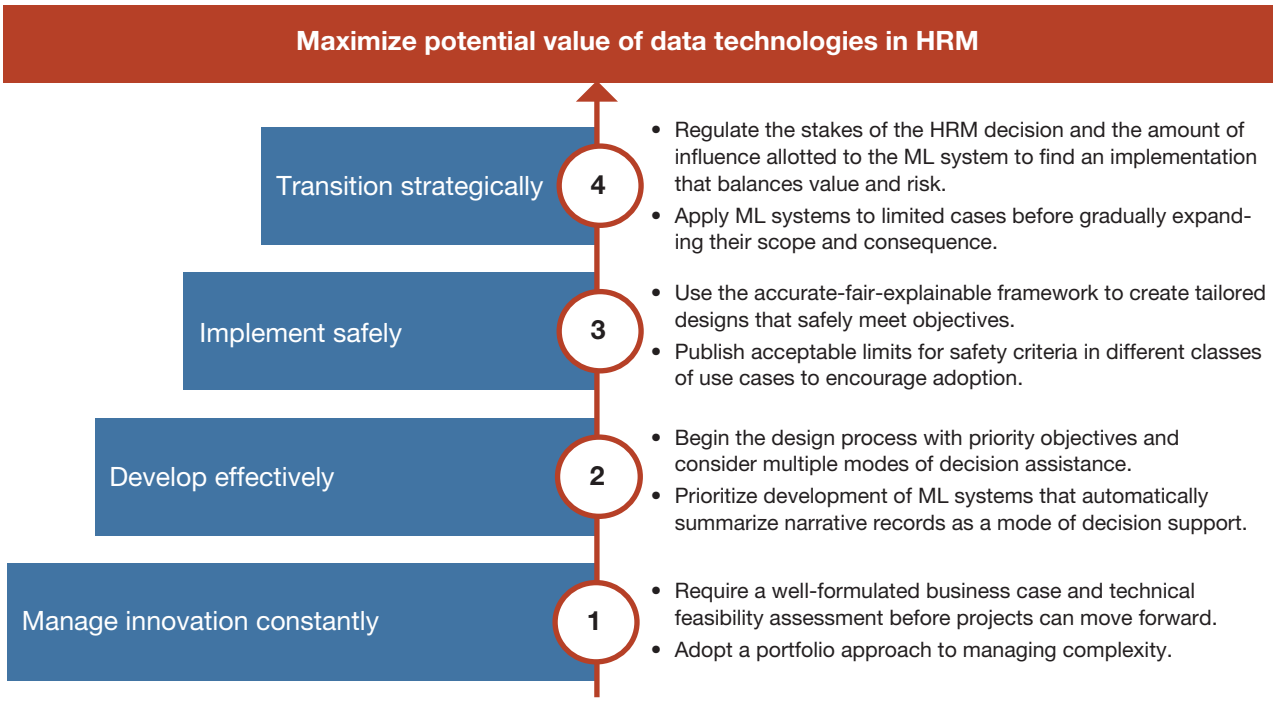
The lesson of the framework is that there is significant flexibility in choosing the implementation of an ML decision-support system that could permit decisionmakers to continue to embrace the agile mantra of "release early, release often."

Conclusion

Successfully leveraging data technologies for value in the HRM domain requires the DAF to broaden its focus beyond particular projects and consider the four pillars that we discuss in this report (Figure 7). Decisionmakers at Headquarters Air Force and in

FIGURE 7

Summary of Recommendations for Leveraging Data Technologies in HRM



key HRM centers must select the right portfolio of projects, execute their development effectively, provide developers with the tools to design safe systems, and open up many transition pathways to encourage broader use of systems at different levels of maturity. In this way, the DAF can overcome adoption challenges and begin to move toward strategic objectives for leveraging data technology to improve HRM decisions.

Notes

[1] For example, researchers from the Berkman Klein Center for Internet and Society at Harvard University identified eight themes from the body of work on artificial intelligence principles: privacy, accountability, safety/security, transparency/explainability, fairness/nondiscrimination, human control, professional responsibility, and promotion of human values (Fjeld et al., 2020).

[2] According to professional standards for personnel selection, the essential principle in evaluating any procedure is to establish that it is job-related by tying it to an important aspect of work behavior. These standards use the term *validity* to refer to the demonstration that a system accurately predicts work behavior and, thus, can be interpreted as intended in the selection process (Society for Industrial Organizational Psychology, 2018).

[3] In the HRM domain, it is vital to think carefully about the appropriateness of the data. With many HRM outcomes, including board selection decisions, there is no objective "ground truth" upon which to base model predictions. In every application, a model's capability will be constrained by the usefulness of the available data. If the data are not closely related to the HRM decisions, the model would not be safe to use.

[4] Similarly, professional standards for personnel selection note that "equitable treatment of all examinees during the selection process" is one potential meaning of *fairness*, while other potential meanings focus on fairness as a lack of bias in measurement (Society for Industrial Organizational Psychology, 2018).

[5] Personnel selection guidelines refer to systematic errors against subgroups as *predictive bias* (Society for Industrial Organizational Psychology, 2018).

[6] For example, we show in Volume 4 (Snoke et al., 2024) that a system that applies this design to the O-5 promotion process, while classifying 25 percent of the records outside the gray zone, would be over 98 percent accurate, with false negative errors (those that incorrectly fail to select worthy officers) occurring at a rate of 0.5 percent. This implementation satisfies the independence fairness criterion, given that it selects women at a slightly higher rate than men. And the method behind the system, logistic regression, is considered in many cases to be inherently interpretable.

References

Asch, Beth J., Trey Miller, and Alessandro Malchiodi, *A New Look at Gender and Minority Differences in Officer Career Progression in the Military*, RAND Corporation, TR-1159-OSD, 2012. As of November 10, 2023: https://www.rand.org/pubs/technical_reports/TR1159.html

Brown, Charles Q., *CSAF Action Orders to Accelerate Change Across Air Force*, Chief of Staff, U.S. Air Force, December 2020, Modification 1, February 7, 2022.

Bryar, Colin, and Bill Carr, "Have We Taken Agile Too Far?" *Harvard Business Review*, 2021.

Budhwar, Pawan, Ashish Malik, M. T. Thedushika De Silva, and Praveena Thevisuthan, "Artificial Intelligence—Challenges and Opportunities for International HRM: A Review and Research Agenda," *International Journal of Human Resource Management*, Vol. 33, No. 6, 2022.

Calkins, Avery, Monique Graham, Claude Messan Setodji, David Schulker, and Matthew Walsh, *Machine Learning–Enabled Recommendations for the Air Force Officer Assignment System:* Volume 5, RAND Corporation, RR-A1745-5, 2024.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*, NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc., and OHRA Verzekeringen en Bank Groep B.V., 2000.

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky, "The State of AI in 2021," McKinsey & Company, December 8, 2021. As of November 10, 2023: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Code of Federal Regulations: Title 29, Subtitle B, Chapter 14, Part 1607.4 (D) (eCFR :: 29 CFR Part 1607—Uniform Guidelines on Employee Selection Procedures (1978)).

DoD—*See* U.S. Department of Defense.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Berkman Klein Center Research Publication No. 2020-1, February 14, 2020. As of April 28, 2022: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

Fountaine, Tim, Brian McCarthy, and Tamim Saleh, "Building the AI-Powered Organization," *Harvard Business Review*, July–August 2019.

Guenole, Nigel, and Sheri Feinzig, *The Business Case for AI in HR: With Insights and Tips on Getting Started*, IBM Corporation, 2018.

Josh Bersin Company, *HR Technology 2021: The Definitive Guide*, 2021. As of November 10, 2023: https://joshbersin.com/research/hr-technology-vendors-and-tools/

Nagji, Bansi, and Geoff Tuff, "Managing Your Innovation Portfolio," *Harvard Business Review*, 2012.

National Academies of Sciences, Engineering, and Medicine, *Strengthening U.S. Air Force Human Capital Management: A Flight Plan for 2020–2030*, National Academies Press, 2020.

Pope, Adrian P., Jaime S. Ide, Daria Mićović, Henry Diaz, David Rosenbluth, Lee Ritholtz, Jason C. Twedt, Thayne T. Walker, Kevin Alcedo, and Daniel Javorsek, "Hierarchical Reinforcement Learning for Air-to-Air Combat," in *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, Institute of Electrical and Electronics Engineers, 2021.

Robson, Sean, Maria C. Lytell, Matthew Walsh, Kimberly Curry Hall, Kirsten M. Keller, Vikram Kilambi, Joshua Snoke, Jonathan W. Welburn, Patrick S. Roberts, Owen Hall, and Louis T. Mariano, *U.S. Air Force Enlisted Classification and Reclassification: Potential Improvements Using Machine Learning and Optimization Models*, RAND Corporation, RR-A284-1, 2022. As of November 10, 2023: https://www.rand.org/pubs/research_reports/RRA284-1.html

Schulker, David, Lisa M. Harrington, Matthew Walsh, Sandra Kay Evans, Irineo Cabreros, Dana Udwin, Anthony Lawrence, Christopher E. Maerzluft, and Claude Messan Setodji, *Developing an Air Force Retention Early Warning System: Concept and Initial Prototype*, RAND Corporation, RR-A545-1, 2021. As of November 10, 2023: https://www.rand.org/pubs/research_reports/RRA545-1.html

Schulker, David, Nelson Lim, Luke J. Mathews, Geoffrey E. Grimm, Anthony Lawrence, and Perry Shameem Firoz, *Can Artificial Intelligence Help Improve Air Force Talent Management? An Exploratory Application*, RAND Corporation, RR-A812-1, 2021. As of November 10, 2023: https://www.rand.org/pubs/research_reports/RRA812-1.html

Schulker, David, Joshua Williams, Cheryl K. Montemayor, Li Ang Zhang, and Matthew Walsh, *The Personnel Records Scoring System:* Vol. 3*, A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking*, RAND Corporation, RR-A1745-3, 2024.

Shore, James, *The Art of Agile Development*, 2nd ed., O'Reilly Media, 2022.

Snoke, Joshua, Matthew Walsh, Joshua Williams, and David Schulker, *Safe Use of Machine Learning for Air Force Human Resource Management:* Vol. 4, *Evaluation Framework and Use Cases*, RAND Corporation, RR-A1745-4, 2024.

Society for Industrial Organizational Psychology, *Principles for the Validation and Use of Personnel Selection Procedures*, 5th ed., American Psychological Association, August 2018.

Tambe, Prasanna, Peter Cappelli, and Valery Yakubovich, "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *California Management Review*, Vol. 61, No. 4, 2019.

U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*, 2018.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," 31st Conference on Neural Information Processing Systems, 2017.

Walsh, Matthew, Sean Robson, Albert A. Robbert, and David Schulker, *Machine Learning in Air Force Human Resource Management:* Vol. 2, *A Framework for Vetting Use Cases with Example Applications*, RAND Corporation, RR-A1745-2, 2024.

Walsh, Matthew, David Schulker, Nelson Lim, Albert A. Robbert, Raymond E. Conley, John S. Crown, and Christopher E. Maerzluft, *Department of the Air Force Officer Talent Management Reforms: Implications for Career Field Health and Demographic Diversity*, RAND Corporation, RR-A556-1, 2021. As of November 10, 2023: https://www.rand.org/pubs/research_reports/RRA556-1.html

Acknowledgments

We thank Gregory Parsons, Director of Plans and Integration, Deputy Chief of Staff for Manpower and Personnel, Headquarters U.S. Air Force (AF/A1X), for his support throughout the project. We thank Col Laura King (AF/A1H) and Doug Boerman (AF/A1X). This research benefited greatly from their input and support. We thank the many RAND colleagues who helped with this work: principally, but not exclusively, Melissa Bauman, Benjamin Gibson, Lisa Harrington, Ignacio Lara, Nelson Lim, and Miriam Matthews. We appreciate the comments from peer reviewers Pete Schirmer, Ginger Groeber, and Larry Hanser.

About This Report

The Department of the Air Force (DAF) has begun to develop and field machine learning systems for myriad mission areas and support functions, including human resource management. Independent experts have suggested that such systems have the potential to accelerate existing decision processes and to enhance decision quality. Further, these systems could enable entirely new decision processes to allow the DAF to utilize its human capital more fully.

This project was conceived to develop decision-support methods and tools to help managers and panel members process and understand performance records. We organized our research tasks broadly around the life cycle of data technology adoption and present our findings as a series of tailored reports on different topics:

- *Machine Learning in Air Force Human Resource Management: Volume 2, A Framework for Vetting Use Cases with Example Applications*, by Matthew Walsh, Sean Robson, Albert A. Robbert, and David Schulker, 2024
- *The Personnel Records Scoring System: Volume 3, A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking*, by David Schulker, Joshua Williams, Cheryl K. Montemayor, Li Ang Zhang, and Matthew Walsh, RR-A1745-3, 2024
- *Safe Use of Machine Learning for Air Force Human Resource Management: Volume 4, Evaluation Framework and Use Cases*, by Joshua Snoke, Matthew Walsh, Joshua Williams, and David Schulker, RR-A1745-4, 2024
- *Machine Learning–Enabled Recommendations for the Air Force Officer Assignment System: Volume 5*, by Avery Calkins, Monique Graham, Claude Messan Setodji, David Schulker, and Matthew Walsh, RR-A1745-5, 2024.

The research reported here was commissioned by the Director of Plans and Integration, Deputy Chief of Staff for Manpower and Personnel, Headquarters U.S. Air Force (AF/A1X) and conducted within the Workforce, Development, and Health Program of RAND Project AIR FORCE as part of a fiscal year 2022 project, "Machine Learning Decision-Support Tools for Talent Management Processes."

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of the RAND Corporation, is the Department of the Air Force's (DAF's) federally funded research and development center for studies and analyses, supporting both the United States Air Force and the United States Space Force. PAF provides the DAF with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Strategy and Doctrine; Force Modernization and Employment; Resource Management; and Workforce, Development, and Health. The research reported here was prepared under contract FA7014-22-D-0001.

Additional information about PAF is available on our website: www.rand.org/paf/

This report documents work originally shared with the DAF on September 13, 2022. The draft report, dated September 2022, was reviewed by formal peer reviewers and DAF subject-matter experts.

For more information on this publication, visit www.rand.org/t/RRA1745-1.

www.rand.org