NPS-AM-26-233

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Artificial Intelligence for Unbiased Acquisition Planning: A Case Study in Strategy and Baseline Development

June 2026

CPT Matthew J. Last, USA

Thesis Advisors: Dr. Robert F. Mortlock, Professor
Dr. Mitchell Friedman, Lecturer

Department of Acquisition, Finance and Manpower

Naval Postgraduate School

Prepared for the Naval Postgraduate School, Monterey, CA 93943

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE AND MANPOWER
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Department of Acquisition, Finance and Manpower at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact the Acquisition Research Program (ARP) via email, arp@nps.edu or at 831-656-3793

Acquisition Research Program
Department of Acquisition, Finance and Manpower
Naval Postgraduate School

ABSTRACT

Defense acquisition programs experience persistent cost growth, schedule delays, and performance shortfalls, with research identifying human cognitive biases in early planning as a contributing factor. This capstone project examines whether artificial intelligence (AI) can reduce cognitive bias in the development of acquisition strategies and acquisition program baselines (APB). Using a comparative case study design, this research administered the Joint Common Missile (JCM) program scenario to eight AI models across 240 runs and compared their outputs against 31 human acquisition professionals using statistical analysis and a five-dimension evaluation rubric. Results indicate AI models triggered optimism bias, anchoring, planning fallacy, and confirmation bias at rates equal to or exceeding humans. Ninety-six percent of AI runs selected the single-step strategy ultimately cancelled in 2004, while 77 percent of humans chose incremental approaches matching the program's successful successor. AI achieved near-zero strategic diversity compared to humans' 97 percent of maximum entropy. Despite these shortcomings, AI showed potential as a structured analytical baseline generator when properly constrained. This research recommends employing AI as decision support rather than decision maker, designing structured frameworks that force AI to highlight independent estimate variance, and expanding the field of behavioral acquisition to study AI decision-making behavior.

THIS PAGE INTENTIONALLY LEFT BLANK

ABOUT THE AUTHOR

Captain Matthew Last is a U.S. Army Acquisition officer with a basic branch of Cyber. He enlisted in 2008 as a Blackhawk Helicopter Repairman and served as an Air Crewmember, Flight Instructor, and Platoon Sergeant with the 2-2 Aviation Regiment in South Korea and the 3-10 General Support Aviation Battalion, 10th Mountain Division, deploying twice in support of Operation Enduring Freedom. Selected for the Army's Green to Gold program, he commissioned as a Cyber officer in May 2018 as a Distinguished Military Graduate of Colorado State University. His assignments include Company Executive Officer with the Cyber Training Battalion at Fort Gordon, GA; Team Lead for the 401 Cyber Combat Support Team and Deputy Task Force Raider Commander for JFHQ-C(AF) at Joint Base San Antonio, TX; and Company Commander, 782d Military Intelligence Battalion, Fort Eisenhower, GA. Upon graduation from the Naval Postgraduate School, he will report to the Capability Program Executive for Intelligence and Spectrum Warfare as the Assistant Product Manager for Product Manager Offensive Cyber Warfare.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

To my wife Jordan, thank you for enduring unsolicited explanations about statistics, and why I needed to run one more batch of prompts through yet another language model at the end of the day. You watched me turn my office into a makeshift AI lab, listened to me argue with myself about Fisher's exact tests, and never once questioned why a defense acquisition student had multiple computers running AI models at all hours. Your support made this possible, and your patience with it was nothing short of extraordinary. To my children Hazel and Nathan, thank you for always understanding that I had another assignment and another question to answer before we could play Daddy Monster.

I would like to thank my advisor, Dr. Robert Mortlock, whose research on behavioral acquisition provided both the intellectual foundation and the dataset that made this study possible. Your patience in helping me refine a vague interest in AI into a focused, defensible research question was invaluable, and your willingness to let me push your earlier work in a new direction speaks to the kind of academic mentorship that makes NPS special.

I would also like to thank my co-advisor, Dr. Mitchell Friedman, for his expertise in cognitive factors that sharpened the theoretical grounding of this research. Your commitment to ensuring my writing met the standard expected of an NPS thesis made the final product significantly stronger than what I could have produced on my own.

THIS PAGE INTENTIONALLY LEFT BLANK

NPS-AM-26-233

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Artificial Intelligence for Unbiased Acquisition Planning: A Case Study in Strategy and Baseline Development

June 2026

CPT Matthew J. Last, USA

Thesis Advisors: Dr. Robert F. Mortlock, Professor
Dr. Mitchell Friedman, Lecturer

Department of Acquisition, Finance and Manpower

Naval Postgraduate School

Prepared for the Naval Postgraduate School, Monterey, CA 93943

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAF	Adaptive Acquisition Framework
ACAT	Acquisition Category
ACBS	Acquisition Category Baseline Summary
AI	Artificial Intelligence
APB	Acquisition Program Baseline
AS	Acquisition Strategy
AUPC	Average Unit Procurement Cost
BBP	Better Buying Power
CAIG	Cost Assessment and Program Evaluation
CDD	Capability Development Document
COTS	Commercial Off-The-Shelf
CTE	Critical Technology Element
DAWIA	Defense Acquisition Workforce Improvement Act
DoD	Department of Defense
EMD	Engineering and Manufacturing Development
FY	Fiscal Year
GAO	Government Accountability Office
ICE	Independent Cost Estimate
IG	Inspector General
IOC	Initial Operational Capability
IPB	Intelligence Preparation of the Battlefield
JAGM	Joint Air-to-Ground Missile
JCM	Joint Common Missile
JROC	Joint Requirements Oversight Council
KBA	Knowledge-Based Acquisition
KPP	Key Performance Parameter
LLM	Large Language Model

MCA	Major Capability Acquisition
MDA	Milestone Decision Authority
MDAP	Major Defense Acquisition Program
ML	Machine Learning
MTA	Middle Tier of Acquisition
NDI	Non-Developmental Item
NPS	Naval Postgraduate School
OIG	Office of Inspector General
PM	Program Manager
POE	Program Office Estimate
POM	Program Objective Memorandum
PPBE	Planning, Programming, Budgeting, and Execution
RDT&E	Research, Development, Test, and Evaluation
RLHF	Reinforcement Learning from Human Feedback
SAE	Service Acquisition Executive
SME	Subject Matter Expert
STO	Science and Technology Objective
SVM	Support Vector Machine
TMRR	Technology Maturation and Risk Reduction
TRL	Technology Readiness Level
USD(A&S)	Under Secretary of Defense for Acquisition and Sustainment
USD(R&E)	Under Secretary of Defense for Research and Engineering
WBS	Work Breakdown Structure
WSARA	Weapon Systems Acquisition Reform Act

I. INTRODUCTION

The U.S. Department of Defense (DoD) is responsible for overseeing a large acquisition system that is both expensive and difficult to manage, and it has forecasted expenditures on the most expensive weapons systems in its portfolio to be over $2.4 trillion by the time all of those purchases have been made (Government Accountability Office [GAO], 2025). As the GAO Comptroller General stated, "our government can no longer afford to invest billions of dollars to develop less than the most advanced technologies" in an environment of rising debt and near-peer threats (GAO, 2025, p. 1). In the face of this background, not to mention an intricate strategic landscape and advancing threats, defense acquisition continues to experience the same three issues: one, cost growth that erodes taxpayer investment; two, schedule delays that extend the expected time to provide warfighters capabilities; and three, performance shortfalls that ultimately undermine readiness in an ever-shifting environment (Wong et al., 2022). The DoD weapon systems' 23rd annual report from the GAO found that the expected time to reach initial operational capability for major defense acquisition programs (MDAPs) now stands at nearly 12 years from program start (GAO, 2025). The GAO also highlights that overall costs for 30 evaluated programs grew by $49.3 billion between 2024 and 2025 (GAO, 2025). The Comptroller General noted that "our findings over my 15 years have grown increasingly dire" and that "DoD weapon systems continue to cost more and take even longer to deliver, notwithstanding recent reforms" (GAO, 2025, p. 1). While conditions and challenges vary from program to program, cost, schedule, and performance remain key elements of every program and shortfalls are not anomalies (Wong et al., 2022).

The sources of underperformance have been a topic of analysis and discussion among oversight organizations, researchers, and analysts for years, and are generally agreed to include technical complexity, funding instability, and industrial base constraints, among others (Fox, 2011; Wong et al., 2022). Researchers have identified a less visible set of causes with roots in human cognitive capabilities and decision-making biases (Mortlock & Dew, 2021; Wong et al., 2022). Mortlock and Dew (2021) found "strong evidence that systemic behavioral biases affected the management and decision-

making within these programs," identifying four biases of relevance: planning fallacy, difficulty in making trade-offs, over-optimism, and recency bias (p. 111). Early planning documents, particularly the Acquisition Strategy (AS) and Acquisition Program Baseline (APB), set the forecast for cost, schedule, and performance that govern a program for years (Drezner & Krop, 1997; DoD, 2021, 2022b). Drezner and Krop (1997) found that "most programs experience events that result in a baseline breach at some point in their life-cycles" (p. xi) and that the baselining process historically applied uniform thresholds "regardless of the fact that some deviations from the baseline are inherently more important than others" (p. xi). Programs have continued to suffer from the same pattern: including flawed early assumptions generate cost growth, schedule delays, and performance shortfalls that reverberate across the program's life (GAO, 2015, 2018, 2020, 2025).

Regarding behavioral issues involved, the field of cognitive psychology has produced a wealth of evidence that humans are prone to systematic errors stemming from reliance on heuristic shortcuts, especially under conditions of high complexity and ambiguity (Tversky & Kahneman, 1974). Tversky and Kahneman (1974) demonstrated that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" and that while "these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (p. 1124). Defense acquisition environments frequently experience these conditions, including information gaps, multiple stakeholders, ambiguous requirements, technological risk, schedule pressure, and resource constraints (GAO, 2018, 2020, 2025; Mortlock & Dew, 2021; Tversky & Kahneman, 1974).

As the DoD continues to experience acquisition challenges, the recent rise of artificial intelligence (AI) and machine learning (ML) offers an opportunity to improve or augment human decision-making (Csaszar et al., 2024; Miedema et al., 2026). AI systems can handle large amounts of information, surface connections across data, and even suggest recommendations independent of things that influence human judgment (Csaszar et al., 2024; Harris, 2020; Narbaev et al., 2024). A 2024 study by the RAND Corporation examined how AI applications are used in military operations and possible human biases in the U.S. Army's Intelligence Preparation of the Battlefield (IPB)

processes (Stebbins et al., 2024). The AI-augmented application demonstrated the ability to surface gaps and support the force, increasing confidence among military officers familiar with the human-centric approach in both the analyst and the process (Stebbins et al., 2024). While this RAND study is suggestive of the potential for AI-enabled decision support for complex military and national security problems, its use in the acquisition context, particularly as it might apply to acquisition strategy and baseline development, points to the need for further study.

A. PROBLEM STATEMENT

Human cognitive biases directly influence two primary planning tools in the DoD acquisition system: ASs and APBs (Kiesling & Chong, 2020; Mortlock & Dew, 2021). Acquisition reform and modernization efforts have not succeeded in eliminating program cost growth, schedule delays, and resulting performance deficits (Fox, 2011; GAO, 2025; Wong et al., 2022). Numerous studies indicate that the reasons for these issues lie in the failure of adequate early program planning and subsequent decision-making processes, specifically those related to developing a programmatic strategy and establishing a baseline (GAO, 2025; Drezner & Krop, 1997; Wong et al., 2022). Additionally, these issues are generally not caused by technical or financial risk factors but rather by cognitive bias, suboptimal judgments, and other forms of path dependency in decision making which include optimism bias, anchoring, confirmation bias, and institutional cultural attachment to legacy approaches or solutions (Hammond et al., 2006; Mortlock & Dew, 2021; Tversky & Kahneman, 1974).

Recent reforms and high-level policy statements increasingly advocate for data-driven acquisition and digital engineering (DoD, 2022a, 2022b; Mortlock & Dew, 2021). The acquisition system does not have standardized mechanisms to mitigate human bias in the most important planning artifacts (DoD, 2022a, 2022b; Mortlock & Dew, 2021). The GAO repeatedly identifies the overly optimistic nature of cost and schedule baselines as the main cause of program under-performance. ASs also provide analyses that favor incumbents and do not critically analyze alternative approaches (GAO, 2015, 2020, 2025). Even Program Managers (PMs) who have extensive experience and training are susceptible to service culture, organizational risk aversion and cognitive biases when

developing foundational program documents (Flyvbjerg et al., 2009; Mortlock & Dew, 2021).

Researchers have applied AI and ML technologies to other defense decision-making contexts over the last several years in areas such as intelligence analysis, and in industries like logistics and maintenance (Ayvaz & Alpay, 2021; Lau et al., 2004; Stebbins et al., 2024). However, few researchers have examined how and whether AI could play a role in acquisition decision-making. In particular, the literature offers little inquiry into known and previously tested algorithms or algorithmically generated acquisition decisions against which to test human decisions using the same data. As a result, the field does not yet know whether AI can help make more accurate, objective, and consistent ASs or APBs, whether AI introduces different and more problematic biases, or how AI recommendations might be evaluated and integrated into formal decision processes (Alon-Barkat & Busuioc, 2023; Harris, 2020). These problems are both practical ones for acquisition management and reflect a knowledge gap in the academic literature on defense acquisition and AI-enabled decision support.

B. RESEARCH QUESTIONS

This research addresses the primary question: "To what extent does AI reduce or mitigate human cognitive bias in the development of ASs and APBs?" Secondarily, this research attempts to answer, "How do AI-generated acquisition decisions differ from those produced by human acquisition professionals when analyzing the same case study data set?" and "What opportunities and limitations arise when integrating AI-driven decision-support tools into early acquisition planning, and how might these tools improve the objectivity and consistency of DoD acquisition outcomes?"

C. RESEARCH METHODOLOGY OVERVIEW

This research uses a comparative case study design with mixed-methods data collection and analysis to explore the degree to which AI might address or overcome human cognitive bias in the context of defense acquisition decision making. The design and implementation allow for a comparison of human and AI-produced acquisition

planning outputs to determine differences in decision quality, analytical rigor, and specific evidence of cognitive bias.

The design centers on a single acquisition case study drawn from an actual program event with established requirements for an AS and APB. This study compares human acquisition practitioner outputs, produced under normal conditions from the case study, against artifacts that AI models generate from the same data. The study provides multiple AI systems with the same inputs to produce a comparable set of outputs. This allows for the comparison of decision-making performance across AI models. A key advantage of this method is that it controls both the human vs. AI variables and the decision-making environment/context.

Primary data for comparison includes human-produced survey results, AI-produced results from normalized input prompts, and scores that human subject matter experts (SME) assign to all artifacts using an objective rubric. The evaluation rubric combines GAO acquisition knowledge best practices with well-established cognitive bias flags from the cognitive and behavioral research literature (GAO, 2018, 2020; Mortlock & Dew, 2021; Tversky & Kahneman, 1974). This analysis compares the two sets of results, human and AI, for evidence of optimism bias, anchoring, legacy preference, other heuristics, as well as measures of analytical rigor, internal consistency, and policy conformance.

The experiment identifies differences in decision quality and bias across AI and human sources to help assess the value-add of AI-enabled acquisition decision-support tools for early planning. The comparative design framework helps identify specific dimensions along which AI recommendations are superior, commensurate, or inferior to human professional judgment.

D. SCOPE AND LIMITATIONS

The primary area of study is early-stage acquisition planning, including the development of ASs and APBs. Decisions regarding acquisition requirements and other items made during the initial stages of an acquisition directly impact long-term program performance (Drezner & Krop, 1997; GAO, 2020). Later reviews, re-baselining, and

refinement of ASs and programs have little practical influence on initial judgments and decisions (Drezner & Krop, 1997). This research uses a single acquisition case study detailed enough to yield human-AI comparison. Study of a single specific example controls for variation between human and AI inputs while retaining validity by using a realistic and grounded scenario description, with both human and AI outputs judged on the same criteria against the same information set and with the same constraints.

Aspects considered in the case study include strategy elements, baseline parameters, cost and schedule assumptions, risk identification and prioritization, and rationale behind the choices. A set of bias constructs interprets potential human-AI differences using the cognitive bias framework that behavioral decision-making researchers have developed and refined over several decades (Gilovich et al., 2002; Kahneman, 2011; Mortlock & Dew, 2021; Tversky & Kahneman, 1974).

Some limitations in this research restrict the applicability of its conclusions. First, the choice of a single case study with small sample size naturally impacts generalizability to other types of programs and to all acquisition categories. Different programs, depending on specific characteristics of the commodity or service in question, the Service component, the acquisition pathway, or the level of existing technology maturity and developmental readiness, could have interactions with both cognitive biases and AI tools that a single example does not represent.

Second, human decisions made in an educational or research setting as an individual do not capture the full range of conditions, including organizational and hierarchy pressures, political influences, power dynamics, or broader stakeholder considerations, as well as time or other resource constraints, that are present in a real-world program environment and may influence bias (Mortlock & Dew, 2021). The simplified research environment necessary for comparison and analysis abstracts from the more complex, realistic acquisition operational context.

Third, different choices in AI model architecture, training data, and prompting scheme or structure produce different outputs and mean that research results do not necessarily generalize to all AI systems or to future AI technology (Csaszar et al., 2024; Miedema et al., 2026). The AI systems themselves might introduce AI-specific biases

related to composition and provenance of training data, technical model architecture choices, or prompt engineering choices (Miedema et al., 2026). This research does not control for or analyze these biases.

Fourth, this research does not evaluate or attempt to mitigate AI-specific risks, such as hallucinations (confident but false or misleading statements), bias in training data, or lack of transparency, because much more rigorous research would be necessary to understand and control for these factors, which remains outside the scope of this work (Miedema et al., 2026). While untrained models used for the study generally sidestep the AI-specific risks, these are still necessarily present in AI to some extent and can occur periodically (Miedema et al., 2026). This research also does not specifically discuss ethical, policy, or legal issues associated with the use of AI to assist with acquisition decisions. These would be a topic for a separate research effort analyzing those specific components of decision-making environment.

Finally, due to the use of AI to develop responses to satisfy this research, portions of this document contain AI generated content that may be flagged by AI checking software. While no AI checking software is 100% correct, it will be noted within this document when AI responses are presented, and which models were used to create the associated response. This allows for transparency in the process and ensures that the research does not inadvertently become disqualified by institutions for AI generated content.

Results from the study are therefore to be considered preliminary and foundational, and the research is intended to be an initial exploration to yield early insights into the potential role of AI to produce more objective, consistent, and reliable acquisition planning outputs, on which subsequent research can build to expand scope and address the limitations as identified in future work.

E. ORGANIZATION OF THE PROJECT

Five chapters structure this project, each building on the previous one to develop and then report on the research study, and to draw conclusions about the possibility and potential utility of using AI to improve acquisition planning by overcoming cognitive

bias. Chapter I provides the introduction to the project, which orients the reader to the research problem, research questions, the methodology, scope and limitations, and the organization of the project. Chapter II provides the background to the study by addressing the defense acquisition environment, including an overview of the DoD acquisition system, the purpose and content of ASs and APBs, and the influence of early acquisition decisions on future program performance; it also summarizes the historical performance of acquisition programs to provide an empirical basis for the research problem. Chapter III presents a review of the relevant literature that develops the theoretical basis for the study; the review also identifies and discusses relevant gaps in the current literature. Chapter IV reports on research, covering methodology, data collection, and analysis. This chapter addresses the case study approach, human and AI data sources, rubric development for evaluation, and comparison results. It compares findings on decision quality, evidence of cognitive bias, and differences in analytical characteristics. Chapter V provides a summary of the findings, responses to the research questions, conclusions about the potential role of AI in acquisition planning, and recommendations for policy and practice and future research. The final section also addresses the broader implications for the DoD acquisition decision-making process and the need for further work in this area.

II. BACKGROUND

Chapter I established that cost growth, schedule delays, and performance shortfalls persist across defense acquisition programs despite decades of reform, and that cognitive biases may represent an underappreciated cause. This chapter provides the contextual foundation necessary to understand where biases enter and why they matter.

A. OVERVIEW OF THE DEFENSE ACQUISITION SYSTEM

Understanding where cognitive biases enter acquisition decision-making first requires an understanding of the system in which those decisions are made. This section provides a structural overview by first establishing the scale and financial magnitude of DoD acquisition, then identifying the key organizations and authorities that govern program decisions, and finally describing the Adaptive Acquisition Framework that defines the pathways and milestones through which programs advance.

1. Size and Scope of DoD Acquisition

DoD Directive 5000.01, "The Defense Acquisition System," states that the Defense Acquisition System helps achieve the National Defense Strategy by developing

> a more lethal force based on U.S. technological innovation and a culture of performance that yields a decisive and sustained U.S. military advantage. The acquisition system will be designed to acquire products and services that satisfy user needs with measurable and timely improvements to mission capability, material readiness, and operational support, at a fair and reasonable price. (DoD, 2022a, p. 4)

In the fiscal year (FY) 2026 Presidential Budget, the DoD requested $384.3 billion (40%) of the total $961.6 billion dollar defense budget request for use in acquisition program funding spread across procurement and research, development, test, and evaluation (RDT&E) appropriations (DoD OIG, 2025). The DoD's "request will fund over 2,049 DoD acquisition programs, projects, and activities" (DoD OIG, 2025, p. 1). As the DoD Inspector General noted, "Robust and continuous acquisition planning is crucial to ensure the DoD's ability to execute the National Defense Strategy and deliver weapon systems with the right capability, at the right time, and at the best cost" (DoD

OIG, 2025, p. 1). The DoD plans to make substantial investments approaching $2.4 trillion to develop and acquire its most costly weapon systems over several years (GAO, 2025). Comptroller General Dodaro emphasized that "the need for smart spending and increased urgency and innovation in DoD's weapon system acquisitions are national imperatives" and that "our government can no longer afford to invest billions of dollars to develop less than the most advanced technologies in an environment of mounting federal debt and ascendent near-peer threats" (GAO, 2025, p. 1). Therefore, the DoD needs robust acquisition planning efforts to ensure it can execute the National Defense Strategy while delivering weapon systems that provide the right capability at the right time in the most cost-effective manner (DoD OIG, 2025).

2. Key Organizations and Authorities

The Defense Acquisition Governance Framework divides authority and accountability across all organizational levels (DoD, 2022a). The Under Secretary of Defense for Acquisition and Sustainment (USD[A&S]) "serves as the MDA for the Materiel Development Decision, Milestone A, the Request for Proposal Release Decision Point for the Engineering and Manufacturing Development Phase, Milestone B, and Milestone C for acquisition category (ACAT) ID programs" (DoD, 2022a, p. 11). The USD(A&S) also "issues and maintains requirements for the content, review, and approval process for ACAT ID acquisition strategies" (DoD, 2022a, p. 11). The Under Secretary of Defense for Research and Engineering (USD[R&E]) "confirms that a materiel solution that addresses the validated need or capability gap for the MDAP is technically feasible and achievable" and "conducts and approves independent technical risk assessments (ITRAs) for ACAT ID Programs" (DoD, 2022a, p. 11). The Service Acquisition Executive (SAE), as the Senior Procurement Executives in their respective Services, are typically the Milestone Decision Authorities (MDAs) for Major Programs in their portfolios (GAO, 2025).

The PM is the focal point for the execution of an acquisition program. In addition to developing an acquisition strategy, PMs are responsible for managing the program's baseline, including defining constraints on time, resources and desired performance (DoD, 2021, 2022b). Mortlock and Dew (2021) highlight that PMs operate at the center

of three systems that exert force in different directions. One, requirements generation is needs-driven and responds to evolving threats and warfighter needs (Mortlock & Dew, 2021). Two, resource allocation is calendar-driven, associated with annual congressional appropriations bills (Mortlock & Dew, 2021). Three, the Adaptive Acquisition Framework (AAF) is event-driven and moves programs forward in accordance with milestone and decision events (Mortlock & Dew, 2021). As they note, "each of these decision support systems is fundamentally driven by different and often contradictory factors" (Mortlock & Dew, 2021, p. 95). PMs must balance these competing systems at every decision point, creating conditions opportune for shortcuts and bias (Mortlock & Dew, 2021; Tversky & Kahneman, 1974).

3. The Adaptive Acquisition Framework

The DoD implemented the AAF at the beginning of 2020 (DoD, 2022a, 2022b). The DoD created six acquisition pathways as shown in Figure 1: Urgent Capability Acquisition, Middle Tier of Acquisition (MTA), Major Capability Acquisition (MCA), Software Acquisition, Defense Business Systems, and Acquisition of Services (DoD, 2022b). The framework emphasized delegation of decision authority as well as tailoring acquisition pathways based on program attributes and managing risks throughout the life cycle (DoD, 2022b). MCA, under DoD Instruction 5000.85, remains the pathway for MDAPs. The MCA pathway consists of the following phases: Materiel Solutions Analysis, Technology Maturation and Risk Reduction (TMRR), Engineering and Manufacturing Development (EMD), Production and Deployment, and Operations and Support (DoD, 2021). Milestone Decisions are used by DoD to determine if a program can move from one Phase to another as well as to set out planning assumptions that will be used throughout the life of the Program (DoD, 2021).

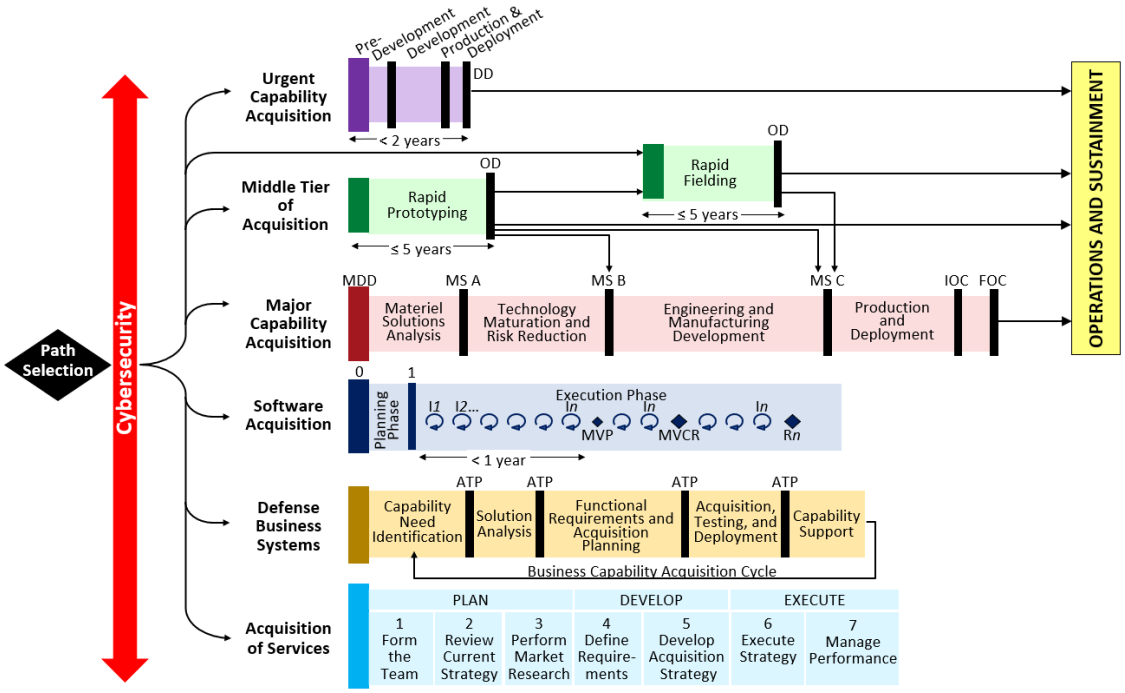

Figure 1. The Adaptive Acquisition Framework. Source: DoD (2020).

B. FOUNDATIONAL ACQUISITION DOCUMENTS

Two documents anchor every major program's trajectory from its earliest planning stages through execution. The following sections describe the AS and APB and why the assumptions embedded within them during early planning exert disproportionate influence over long-term program outcomes.

1. The Acquisition Strategy

The AS documents the overall approach to meet program objectives and includes the business strategy, technical strategy, and support strategy into one integrated program plan (DoD, 2022b). The AS will define the program's overall acquisition approach and incorporates the acquisition pathway, contract strategy, competitive strategy, source selection plans, testing and evaluation, as well as integrating this program or system with other programs or systems (DoD, 2022b). The PM creates this strategy plan, and the MDA approves it at each milestone decision (DoD, 2022b).

The AS should consider and satisfy various statutory and policy requirements such as those associated with competition, intellectual property rights, modular open

system approaches, cybersecurity, sustainment, and small business (DoD, 2022b). Choosing the right strategy early in the program can be challenging, especially when there is great uncertainty about the future requirements and how those will be fulfilled by available technologies, commercial or development, and under what conditions they will operate commercially (Mortlock, 2020). PMs must make decisions about the future with significant uncertainty surrounding the technology available, what the marketplace will look like and what capabilities are needed (Mortlock, 2020). Moreover, cognitive biases most influence human decision-making during situations that involve significant uncertainty and complex conditions, according to Tversky and Kahneman (1974).

Decisions made during strategy development affect the entirety of the program (Drezner & Krop, 1997). Every decision regarding contract type or technical risk acceptance and competition structure narrows down future options (Drezner & Krop, 1997; Mortlock, 2020). Mortlock and Dew (2021) observe that "due to the inherent complexity of the development, procurement, and fielding of sophisticated weapons systems that are required to operate reliably in challenging military environments, acquisition programs often fail to deliver required performance capabilities within cost and schedule constraints" (p. 95). The GAO Cost Estimating and Assessment Guide confirms that "many cost estimating challenges can be traced to over-optimism" in the technical baselines provided by program offices (GAO, 2020, p. 10).

2. The APB

The APB is the formal contract that outlines the PM's commitment to the MDA regarding the cost, time and level of performance of a program (DoD, 2021). The APB generally begins upon entry to EMD after Milestone B. At this point it defines two levels of value; Threshold, which are minimum acceptable values and Objective, or desired values for Cost, Schedule and some Key Performance Parameters (KPPs) (DoD, 2021). KPPs in the Capability Development Document (CDD) then flow into the APB providing strong linkage between requirements and baseline commitments (Drezner & Krop, 1997).

Evidence from past programs demonstrates their reliance upon their baselines due to path dependency providing a rationale for this reliance (Drezner & Krop, 1997). Once set, baselines guide decisions about resource allocations in the Planning, Programming,

Budgeting, and Execution (PPBE) process, language used in congressional justifications, and expectations of stakeholders. Programs face significant institutional expenses, both funding and programmatic, when they alter their baselines (Drezner & Krop, 1997). Once approved, alterations to the baseline trigger new reporting thresholds and additional oversight organizations (Drezner & Krop, 1997). Managers have incentive to adjust estimates up when requesting approval of new programs, to ensure "selling" the program (Drezner & Krop, 1997). The GAO (2015) found that PMs face structural incentives to design acquisition strategies around milestone approval rather than long-term executability and capability delivery. Managers also have incentive to not update their baseline if it becomes unrealistic to avoid highlighting problems (Drezner & Krop, 1997). The GAO Cost Estimating and Assessment Guide documents this pattern directly, stating "History has shown a clear pattern of higher cost estimates the further away from the program office that the ICE is created. This is because the ICE team is more objective and less prone to accept optimistic assumptions" (GAO, 2020, p. 126).

The Nunn-McCurdy provisions are cost growth threshold levels for acquisitions with statutory levels for notification (GAO, 2020). When an acquisition program experiences significant breach (15%), the program is required to notify (Level I), when it has critical breach (25%) the program notifies (Level II) (GAO, 2020). Programs experiencing a breach will be required to document all the cost increases experienced by the program and either restructure or face termination (Weapon Systems Acquisition Reform Act of 2009, Pub. L. No. 111–23, § 206).

3. Why Early Decisions Matter So Much

Decisions made during initial acquisition planning stages become increasingly difficult to adjust once an acquisition program matures (Drezner & Krop, 1997; GAO, 2020). Design choices, contract obligations, and funding decisions constrain future options. The headwaters of each program narrow logical pathways forward and limit available options (Drezner & Krop, 1997). The GAO Cost Estimating and Assessment Guide makes clear that when a program begins with an unrealistic baseline, it almost certainly experiences cost growth and schedule delays that become increasingly difficult to address as the program progresses (GAO, 2020). As the GAO states explicitly, "if the

baseline is not based on a reliable cost estimate or does not reflect the approved work, the program is at risk for cost overruns, missed deadlines, and shortfalls in performance" (GAO, 2020, p. 210).

The creation of baseline documents and strategy planning serves as groundwork for subsequent program activities and errors or biases that enter during document development compound during program execution. If the APB isn't realistic when approved at Milestone B, the program will likely experience cost growth, schedule slips, and not deliver required capabilities (Drezner & Krop, 1997; GAO, 2020). Multiple studies (Drezner & Krop, 1997; GAO, 2018, 2020, 2025; Wong et al., 2022) have confirmed that the cost and schedule results that programs ultimately achieve are heavily dependent on assumptions and estimates made during early planning documents. The GAO's review of space system acquisitions illustrates this dynamic: in five of six programs reviewed, "program officials and cost estimators assumed when cost estimates were developed that critical technologies would be mature and available. They made this assumption even though the programs had begun without complete understanding of how long they would run or how much it would cost" (GAO, 2020, p. 11, citing GAO-07-96). The NPOESS satellite program was subsequently "beset by significant cost increases and schedule delays, partly because of technical problems" stemming from those early assumptions (GAO, 2020, p. 11). This leverage points to early acquisition planning decisions as the phase where intervention can deliver the biggest returns on time and effort spent.

AS and APB development is the foundation for this research precisely because of the leverage points discussed. If AI-enabled decision support tools can reduce bias in the development of these documents, program execution will amplify those improvements. Small improvements in the objectivity and accuracy of early program plans can have significant downstream impacts on the program's chances of operating within the APB cost, schedule, and performance constraints.

C. HISTORICAL PERFORMANCE CHALLENGES IN DEFENSE ACQUISITION

The following sections examine three persistent dimensions of underperformance, cost growth, schedule delays, and knowledge gaps at milestone decision points, that collectively define the scale of the problem this research seeks to address.

1. Cost Growth Trends

According to the GAO, acquisition has been on the list of high-risk programs since 1990 (GAO, 2025). Programs on the list for more than 30 years demonstrate enduring challenges that persist despite reform efforts (GAO, 2025). While the degree of cost growth fluctuates across different studies, the RAND Corporation found that "dollar-weighted average development cost growth" across 46 programs spanning three decades "was almost 60 percent relative to the Milestone B…estimate" (Wong et al., 2022, p. 23). These escalations are not isolated incidents or the result of unexpected events, they are representative of potential systemic issues (Wong et al., 2022). The Ford-class aircraft carrier program provides a recent illustration. GAO found that "cost growth for the lead ship was driven by challenges with technology development, design, and construction, compounded by an optimistic budget estimate. Instead of learning from the mistakes of CVN 78, the Navy developed an estimate for CVN 79 that assumed a reduction in labor hours needed to construct the ship that was unprecedented in the past 50 years of aircraft carrier construction" (GAO, 2020, p. 26, citing GAO-17-575). The 2025 Weapon Systems Annual Assessment report from the GAO provides a recent observation of this steady problem. Of the 30 MDAPs that were on last year's report, total estimated costs increased by $49.3 billion (GAO, 2025). The Air Force Sentinel intercontinental ballistic missile program drove $36 billion (73%) of that growth (GAO, 2025). Sentinel is an extreme case where the actual costs are far from initial estimates, but it demonstrates how far off initial estimates can be (GAO, 2025).

Many issues can lead to cost growth such as technical challenges, evolving requirements, funding instability, labor/material cost growth, testing failures, and integration difficulties (GAO, 2020; Wong et al., 2022). However, all these items can trace their root cause to estimates that did not incorporate enough realism to include risk

(GAO, 2020; Wong et al., 2022). This pattern is not confined to defense. In a study of major projects across 20 countries, Flyvbjerg et al. (2009) reported that "nine out of ten projects had cost overruns" (p. 171), with overruns above 100 percent not uncommon. Their study suggests the problem is systemic across large public and private ventures (Flyvbjerg et al., 2009). According to the GAO Cost Estimating and Assessment Guide, data bias and estimating practice are primary contributors for cost growth, and resource constraints to the estimating process add to those challenges (GAO, 2020). The GAO notes that "limited program funding and available time often hinder broad participation in cost estimation processes and force the analyst (or cost team) to reduce the extent to which trade-off, sensitivity, and even uncertainty analyses are performed" (GAO, 2020, p. 10).

2. Schedule Delays

Schedule delays have proven to be just as persistent as cost growth across programs (GAO, 2025; Wong et al., 2022). The 2025 Annual Acquisition delays report from the GAO revealed that MDAPs took 18 months more to deliver initial capability starting from program inception compared to the previous year (GAO, 2025). The new average was now approaching 12 years (GAO, 2025). This includes programs that started under the MTA pathways established to deliver capability faster (GAO, 2025). Programs following pathways intended to provide greater speed are still late, which implies that the factors causing delays are broader than any one program's process.

Delays have repercussions throughout the defense enterprise. Development takes longer so it costs more. Longer lead times require paying for personnel and contractor support longer than originally planned (Wong et al., 2022). By the time long development programs are fielded they can be nearing obsolescence, especially in the world of software technology (GAO, 2018; Wong et al., 2022). Warfighters must wait longer for new capabilities, and programs have less time to integrate with other programs and systems as interfaces change (GAO, 2025). PMs and planners discount past performance of similar programs underestimating the effects that similar risks and issues had on past programs' progress. They overestimate their ability to perform better under the same circumstances (Mortlock & Dew, 2021; Kahneman & Lovallo, 1993).

Researchers call this tendency toward unrealistically optimistic schedule completion the planning fallacy, and they have studied it in laboratory settings and in the field (Kahneman & Tversky, 1979; Kahneman & Lovallo, 1993). Defense acquisition programs seem prime candidates for falling into the planning fallacy trap considering their inherent complexity, lengthy timelines, and enterprise incentives (Mortlock & Dew, 2021). Mortlock and Dew (2021) elaborate that "the planning fallacy creates biased expectations that mask a control gap that will exact a price over the course of most programs" (p. 98). Furthermore, "planning processes inevitably understate unpredictable events that will disrupt and delay the plan" because plans are always subject to the potential of "'unknown unknowns' to intervene in what is otherwise carefully manicured expectations" (p. 98). The consequences are direct. As the GAO warns, "these delays put the warfighter at risk of receiving weapon systems that do not deliver needed capabilities" (GAO, 2025, p. 1).

3. Knowledge Gaps at Decision Points

A major theme of the GAO's assessments has been programs advancing through milestone decision points without achieving the knowledge required to have high confidence in cost and schedule estimates (GAO, 2018). The language from the 2018 assessment is clear that the DoD did not know enough about its major programs for leadership to have high confidence in eventual outcomes (GAO, 2018). Programs were advancing before demonstrating knowledge-based practices of technology maturity prior to starting system development, design stability prior to starting production, and proven production processes prior to starting full-rate manufacturing (GAO, 2018), as shown in Figure 2. As the GAO concluded in 2018, despite reforms, programs "continue to proceed without the key knowledge essential to good acquisition outcomes" (GAO, 2018, Highlights). Moreover, programs initiated before 2010 had "experienced cost growth of 62.4 percent, or $542.1 billion, and schedule delays totaling 35 months on average since those programs developed their original estimates" (GAO, 2018, p. 28).The 2025 GAO assessment also found that planned future major weapon acquisitions indicated they would be incorporating leading practices at levels that were at or below the levels of current MDAPs and MTA programs (GAO, 2025). After multiple decades of

experimentation and documentation of consequences flowing from these types of knowledge gaps the defense acquisition system has not yet systematically learned from its mistakes, it requires evaluation if there is something more fundamental than just process design at issue (GAO, 2025).

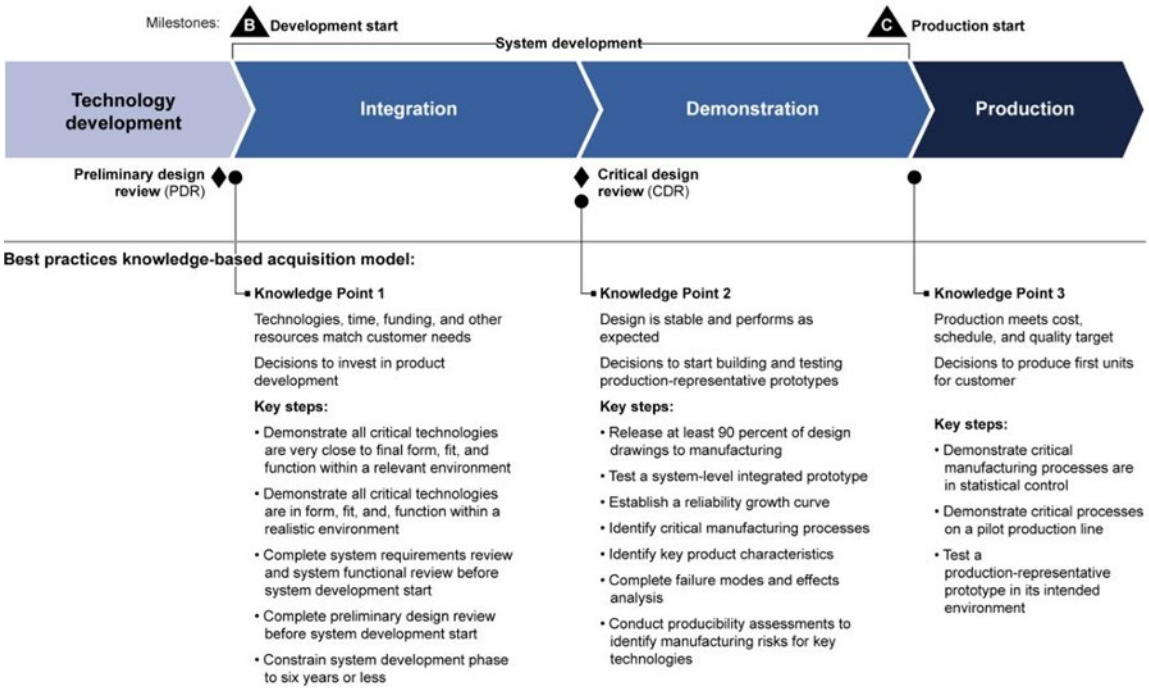

Figure 2. Defense Acquisition Cycle and GAO-Identified Knowledge Points. Source: GAO (2018).

If programs reach decision points without high confidence in knowledge about technology maturity, design stability, or production readiness, decision makers must rely on judgment to bridge gaps in objectively available knowledge (GAO, 2018). This situation represents an ideal case where heuristics together with biases will likely influence and distort decision making (Tversky & Kahneman, 1974). A strategic choice to move forward through milestone-decision points despite being aware of a lack of information may be due to cognitive bias such as overconfidence and optimism for which there is no rational basis (Moore & Healy, 2008; Mortlock & Dew, 2021). As Tversky and Kahneman (1974) demonstrated, "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (p. 1124). The GAO provides direct

evidence of this phenomenon: "it is well known that program advocates tend to underestimate the technical challenges facing the development of a new system" and "technology maturity assumptions also tend to be optimistic" (GAO, 2020, p. 78). Programs that relied on immature technologies at the start "represent a significant challenge and add a high degree of risk to a program's schedule and cost" (GAO, 2020, p. 78).

D. HISTORY OF ACQUISITION REFORM

The cost growth, schedule delays, and knowledge gaps documented in the preceding section are not problems the DoD has ignored, they have prompted over six decades of reform initiatives aimed at improving acquisition outcomes. This section traces historical reform from its origins through major legislative milestones and concludes by examining why the challenges persist despite sustained corrective effort.

1. Background of Defense Acquisition Reform

Secretary of Defense Robert McNamara's implementation of systems analysis and the Planning Programming Budgeting System launched the modern reform era in the 1960s (Fox, 2011). Programs such as the C-5A transport aircraft experienced tremendous cost growth despite then-cutting edge analytical processes, showing that even systems analysis couldn't save acquisition from itself (Fox, 2011). From troubled programs such as the C-5A through the 1970s and early 1980s, procurement scandals dominated headlines with excessively priced spare parts fueling additional rounds of reform, including the Commission on Government Procurement (1969–1972) and the Office of Federal Procurement Policy Act of 1974 (Fox, 2011). Still, the same problems of cost growth and schedule delay persisted (Fox, 2011).

Recommendations from the Packard Commission (1986) had greater bearing on cost and schedule than previous commissions (Fox, 2011). These recommendations included granting greater authority to PMs, improving reliance on prototyping, stabilizing programs, and improving the acquisition workforce through incentives and training (Fox, 2011). Congressional legislation implemented many of these recommendations (Fox, 2011). Despite these reforms (or perhaps because of how they were implemented), in

1990, the GAO added DoD weapon systems acquisition to its High-Risk List where it has remained for more than 35 years (GAO, 2025). The GAO notes that "high-risk lists identify areas of government that have a combination of challenges and opportunities for improvement, and programs that have been on the list for more than 30 years show long-standing, systemic challenges" (GAO, 2025, Highlights).

2. Legislative Reform Targeting Cost and Schedule

Federal legislation began more directly linking policy to cost and schedule successes and failures (Fox, 2011). The Defense Acquisition Workforce Improvement Act (DAWIA; Pub. L. No. 101–510, Title XII, 1990) brought a new level of professionalism to the acquisition field, establishing certification requirements and career development standards for acquisition workforce members (Fox, 2011). Two laws that followed further defined congressional intent that agencies should meet cost, performance, and schedule targets on major programs: the Federal Acquisition Streamlining Act of 1994 (FASA; Pub. L. No. 103–355) and the Clinger-Cohen Act of 1996 (Pub. L. No. 104–106, Division D), originally enacted as the Federal Acquisition Reform Act.

Congress took action by directly targeting cost and schedule issues with the 2009 Weapon Systems Acquisition Reform Act (WSARA; Pub. L. No. 111–23). WSARA required programs to demonstrate maturity of key technologies before entering system development and reformed how Nunn-McCurdy breaches of cost growth were reported (Pub. L. No. 111–23, §§ 104, 206). Under the act, if program costs grow by 15% (significant breach) or 25% (critical breach), the breach must be reported to Congress and the Secretary of the concerned service must consider termination for critical breaches (Pub. L. No. 111–23, § 206). From 2010 onward, a series of initiatives called Better Buying Power (BBP) placed emphasis on affordability analyses and exerting cost control (GAO, 2018; Wong et al., 2022). The GAO's 2018 assessment quantified the impact, finding that it "compared the cost growth of weapon systems development programs for a 5-year period after WSARA's implementation to the 10-year period prior to the act—estimating about a 75 percent, or $36.0 billion, reduction in the rate of development cost

growth" while cautioning that "it is not certain that the act's implementation was the sole reason for this reduction in cost growth" (GAO, 2018, p. 24).

3. Persistence of Challenges Despite Reform

Notwithstanding reforms over the last several decades, program cost growth and schedule slips continue (GAO, 2025; Wong et al., 2022). As Comptroller General Gene Dodaro said about acquisition reform in recent years: "Our findings over my 15 years have grown increasingly dire. DoD weapon systems continue to cost more and take even longer to deliver, notwithstanding recent reforms" (GAO, 2025, p. 1). "The expected time for MDAPs to provide even an initial capability increased this year by 18 months, up to almost 12 years from the program's start—an average that includes MDAPs that began as MTAs" (GAO, 2025, Highlights). On cost, "combined total estimates increased by $49.3 billion for 30 MDAPs also included in last year's report" and "the Air Force's Sentinel missile program accounted for over $36 billion (73 percent) of this increase" (GAO, 2025, Highlights) after experiencing a critical Nunn-McCurdy breach.

After reviewing sixteen reports published from 2020 through 2025 on acquisition issues, the DoD Inspector General (IG) summarized three recurring lessons learned: "(1) develop effective performance requirements, (2) plan and execute adequate test and evaluation procedures, and (3) establish and consistently follow DoD acquisition policy" (DoD OIG, 2025, p. 1). The Packard Commission made two of those three recommendations in 1986 (Fox, 2011).

E. SUMMARY

This chapter laid the foundation for this study by outlining the environment of defense acquisition, introducing the planning documents at the center of this study, describing the long-standing cost, schedule, and performance problems motivating this research, and highlighting reform efforts designed to address them.

The DoD oversees the largest acquisition enterprise in the world and plans to invest nearly $2.4 trillion in its highest-priced weapon programs (GAO, 2025). Every major program begins its life with two planning documents that chart its course forward and establish program assumptions related to cost, schedule, and performance: the AS

and APB (Drezner & Krop, 1997; DoD, 2021, 2022b). Estimates made at program initiation create path dependencies that last for years, if not decades (Drezner & Krop, 1997; GAO, 2020).

When initial planning documents are founded on faulty assumptions, their impacts are realized throughout the life cycle of the program (GAO, 2020). Decades of program data prove that this is exactly what happens (Wong et al., 2022). Since 1990, the GAO has kept defense acquisition on its High-Risk List and in their 2025 report, they found that costs for 30 major programs increased by $49.3 billion in just one year (GAO, 2025). Today, the average time it takes to deliver initial capability is nearly 12 years (GAO, 2025).

Cost growth, schedule slips, and knowledge-truth gaps at milestone decisions have persisted for 35-plus years of reforms from the Packard Commission through WSARA, BBP, and the AAF (DoD OIG, 2025; Fox, 2011; GAO, 2018, 2025). Programs across administrations, reform initiatives, and budget environments are regularly stuck in the same patterns of ineffective decision-making (Wong et al., 2022). If the problem is not with the processes themselves, it must be with the people utilizing them.

Chapter III looks at the body of research from academia to help answer the question. The chapter uses this research to identify how cognitive biases and heuristics relate to defense acquisition programs, provide examples of specific research that has identified cognitive biases within those programs; and highlight current research that is using AI to improve or enhance human decision making. The literature provided in this chapter provides a foundational basis for the central hypothesis of this study: AI-based decision support can reduce or eliminate the cognitive biases that exist in the acquisition workforce.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LITERATURE REVIEW

Chapter II established the scale of the defense acquisition enterprise, the foundational documents that govern program planning, and the persistent challenges that decades of reform have failed to resolve. This chapter examines the scholarly literature across three domains that converge on the central hypothesis of this research: the cognitive psychology research establishing how heuristics and biases systematically distort human judgment, the defense-specific studies documenting those biases within acquisition programs, and the emerging body of work on AI-enabled decision support as a potential countermeasure. The chapter concludes by identifying gaps in the existing research that this study is designed to address.

A. FOUNDATIONS OF HEURISTICS AND BIAS

Decades of research in cognitive psychology has uncovered systematic violations of normative models of judgment and choice under uncertainty and complexity. Psychologists Amos Tversky and Daniel Kahneman pioneered the study of cognitive biases in the 1970s which led to the discovery of dozens of biases that researchers have replicated across many areas (Gilovich et al., 2002). Tversky and Kahneman (1974) demonstrated that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (p. 1124).

Herbert Simon established the conceptual foundations for these studies. His Nobel Prize–winning work argued that "the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man" (Simon, 1957, p. 241). Decision makers respond to these limitations by satisficing—Simon's term for the observation that "organisms adapt well enough to 'satisfice'; they do not, in general, 'optimize'" (Simon, 1957, p. 261). Building upon Simon's work, Tversky and Kahneman's (1974) heuristics and biases program transitioned psychology from broad assertions of cognitive limitation to precise descriptions of how those

limitations lead to systematic and directional error. Since then, the literature of heuristics and biases has generated thousands of scientific articles and affected scholarship in economics, law, medicine, management, and political science (Gilovich et al., 2002).

Researchers have studied many heuristic strategies, but there are three that are applicable to acquisition decisions (Mortlock & Dew, 2021). Tversky and Kahneman (1974) describe how people estimate likelihood based on perceived representativeness. They also discuss how individuals anchor on specific starting points when making predictions and failing to adjust away from them enough (Tversky & Kahneman, 1974). Years later, Kahneman (2011) synthesized behavioral research and popularized the dual-process theory of cognition. Kahneman (2011) defines the two systems precisely: "System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations" (p. 20). Kahneman (2011) explains that "although System 2 believes itself to be where the action is, the automatic System 1 is the hero of the book" because System 1 "effortlessly originat[es] impressions and feelings that are the main sources of the explicit beliefs and deliberate choices of System 2" (p. 21). This statement implies that even if decision makers have the best intentions, all the data in the world, and powerful decision-making tools their judgment is still susceptible to cognitive biases they aren't aware of because System 1 thinking cannot be turned off (Kahneman, 2011).

Kahneman and Frederick (2002) later revised their list of System 1 heuristics to account for the fact that these mental shortcuts tend to take the form of attribute substitution. As Kahneman and Frederick (2002) explained, "when confronted with a difficult question people often answer an easier one instead, usually without being aware of the substitution" (p. 53). For instance, if decision makers are asked to answer a question about probability, they may substitute it with a judgment about similarity or ease of recall instead (Kahneman & Frederick, 2002). Since this replacement happens unconsciously, people are unaware that they have substituted one question for another (Kahneman & Frederick, 2002). Thus, anchoring effects are so pervasive because anchoring values are easy mental shortcuts that people do not sufficiently correct away from, even when the anchor is random and seemingly irrelevant (Tversky & Kahneman,

1974). Ease of recall stems from the availability heuristic (Tversky & Kahneman, 1974). As Evans (1989) explains, Kahneman and Tversky proposed "that people make statistical inferences and judgements by the use of heuristics" where a heuristic is "a short-cut and essentially simple 'rule of thumb'" and that "such biases are likely to be prevalent in real life and expert decision making" (p. 14).

Overconfidence is one of the most pervasive biases in decision making under uncertainty (Moore & Healy, 2008). As stated earlier, when provided with blank confidence intervals, individuals are too confident in their estimates. Tversky and Kahneman (1974) reported that when they asked individuals to provide confidence intervals, they were 98 percent confident in, the "true values fell outside the stated range approximately 30% of the time" (p. 1129) instead of the expected 2% (Tversky & Kahneman, 1974). Researchers have found this bias across both novice and experienced estimators, and it persists even with financial incentives for accuracy (Moore & Healy, 2008).

Moore and Healy (2008) identified three distinct types of overconfidence. The first, overestimation, is "the overestimation of one's actual ability, performance, level of control, or chance of success" (p. 502). The second, overplacement, "occurs when people believe themselves to be better than others" (Moore & Healy, 2008, p. 502). The third, overprecision, is "excessive certainty regarding the accuracy of one's beliefs" (Moore & Healy, 2008, p. 502). Of the three, Moore and Healy (2008) found that "overprecision appears to be more persistent than either of the other 2 types of overconfidence" (p. 502), making it relevant to defense acquisition estimates. Overconfidence is also closely related to the planning fallacy. This occurs when individuals underestimate the factors involved in completing future actions (Kahneman & Lovallo, 1993). Kahneman and Lovallo (1993) articulated the underlying mechanism, arguing that "decision makers have a strong tendency to consider problems as unique" and therefore "neglect the statistics of the past in evaluating current plans" (p. 17). Kahneman (2011) later coined the term planning fallacy to describe "plans and forecasts that are unrealistically close to best-case scenarios" that "could be improved by consulting the statistics of similar cases" (Part III, The Planning Fallacy, para. 1). The mechanism is straightforward: "people who have information about the individual case rarely feel the need to know the statistics of the

class to which the case belongs" and as a result, "in the competition with the inside view, the outside view doesn't stand a chance" (Kahneman, 2011, Part III, Drawn to the Inside View, para. 6–7). Kahneman (2011) warns that "the authors of unrealistic plans are often driven by the desire to get the plan approved—whether by their superiors or by a client—supported by the knowledge that projects are rarely abandoned unfinished merely because of overruns in costs or completion times" (Part III, The Planning Fallacy, para. 7). Consequently, "the greatest responsibility for avoiding the planning fallacy lies with the decision makers who approve the plan. If they do not recognize the need for an outside view, they commit a planning fallacy" (Kahneman, 2011, Part III, The Planning Fallacy, para. 7; Kahneman & Lovallo, 1993). Kahneman et al. (2021) more recently expanded on bias as a sole source of error and introduced the importance of noise, which they defined as "unwanted variability in judgments that should ideally be identical" (Introduction). Noise produces "rampant injustice, high economic costs, and errors of many kinds" (Kahneman et al., 2021, Introduction) across fields from criminal sentencing to insurance underwriting. As the authors summarized, "wherever there is judgment, there is noise—and more of it than you think" (Kahneman et al., 2021, Introduction).

Framing effects also cause bias. An example provided by Tversky and Kahneman (1981) demonstrates how changing the frame of a decision problem can cause people to go from being risk averse to risk seeking. Tversky and Kahneman (1981) found a consistent pattern, noting "choices involving gains are often risk averse and choices involving loss are often risk taking" (p. 453). An example of these phenomena in acquisition would be how an analyst frames the baseline estimate as something to achieve (gain) or something to avoid (loss) systemically changing how people perceive risk in the acquisition process (Tversky & Kahneman, 1981). Prospect theory provides the formal basis for framing effects, which holds that "outcomes are expressed as positive or negative deviations (gains or losses) from a neutral reference outcome" (Tversky & Kahneman, 1981, p. 454). Additionally, Tversky and Kahneman (1981) observe that "the response to losses is more extreme than the response to gains" (p. 454).

In addition to framing, several other biases apply to complex decisions in organizations: Hammond et al. (2006) describe the sunk-cost trap, in which past investments of time or money that are rationally irrelevant to present decisions

nonetheless exert psychological pressure that leads decision makers to continue commitments they would not otherwise choose. Hammond et al. (2006) also describes confirming-evidence bias as a tendency that "leads us to seek out information that supports our existing instinct or point of view while avoiding information that contradicts it" (p. 123). Evans (1989) traces this to Kahneman and Tversky's (1979) broader finding that heuristic-based reasoning produces "a large range of errors and biases on many experimental tasks" (p. 14).

In short, the culmination of research spans decades and counts many cognitive biases has taught psychologists that human judgment is flawed and even trained experts with access to data and decision aids suffer from these biases. We cannot turn off our fast-thinking System 1 and slow-thinking System 2 is not perfect either, being particularly resource intensive, ultimately leading to a consistent vulnerability in decision making (Kahneman, 2011). Cognitive limitations continue to affect decision makers working in uncertain and complex environments, especially present within the defense acquisition domain.

B. DEFENSE-SPECIFIC STUDIES OF COGNITIVE BIAS

Research on cognitive biases in defense acquisition has recently attracted more attention. Mortlock and Dew (2021) investigated behavioral acquisition and analyses three large programs within defense acquisition. The focus here is on the Enhanced Combat Helmet, Joint Common Missile, and Ground Combat Vehicle programs. Mortlock and Dew (2021) found "strong evidence that systemic behavioral biases affected the management and decision-making within these programs" and identified four biases of particular relevance: "planning fallacy, difficulty in making trade-offs, over-optimism, and recency bias" (p. 111). The planning fallacy speaks to program management's unrealistic optimism, which many studies have documented (Kahneman & Lovallo, 1993; Mortlock & Dew, 2021). Essentially, planning fallacy's main argument is that planning processes themselves (independent of other factors) cause managers to believe things about their programs that lead to forecasts that are too optimistic (Mortlock & Dew, 2021). Mortlock and Dew (2021) elaborate that "planning processes lead managers to build an 'inside view' of a project with detailed designs for the

implementation of the project" which enhances "managers' perceptions of control over the project or program. Thus, they become more confident in the success of their plans" (p. 97, citing Kahneman & Lovallo, 1993). In his 2020 case study Mortlock found empirical support for this dynamic. When surveyed, 77% of acquisition professionals recommended an incremental approach over the single-step strategy. However, only 45% of respondents indicated that they would maintain the previously approved cost and schedule. Most respondents selected the option to eliminate or modify all three elements of the triple constraint (cost, schedule, and performance) but did not select to maintain two constraints while relaxing one. It appears that the acquisition professionals had difficulty prioritizing between the three constraints despite stating that there was a need for incrementality (Mortlock, 2020). This inside view dynamic is not unique to defense. Flyvbjerg et al. (2009) found the same pattern in infrastructure megaprojects. They noted that executives adopt an inside view "by focusing tightly on the case at hand" and "constructing scenarios of future progress," rather than consulting the outcomes of comparable past projects (Flyvbjerg et al., 2009, p. 173)

The inability to make trade-offs stems from how difficult it cognitively is to make a direct comparison between two options when they differ in more than one dimension. When programs have multiple requirements that cannot be easily weighed against each other, the decision makers may try to meet all the requirements instead of deciding which ones are most important. This tendency contributes to programs growing too big and complex because decision makers made no trade-offs to simplify them (Mortlock & Dew, 2021). Mortlock and Dew (2021) note that for JCM, "a single-step acquisition strategy to deliver all required capabilities was eventually canceled and the warfighter received no capability. Had an incremental development approach like the subsequent JAGM acquisition strategy been adopted initially, the warfighter could have received improved capability more than a decade sooner" (p. 107). Mortlock (2020) had previously studied this same challenge through the lens of acquisition strategy formulation, finding that policy may need to dictate if programs do not use incremental approaches, they need to provide justification for their decision. Mortlock (2020) suggested that the Defense Acquisition System should break the rigid triple constraint of cost, schedule, and performance. He argued that when all three are fixed simultaneously, programs face an

unnecessarily high risk of failure (Mortlock, 2020). As Mortlock (2020) concluded, the current system "incentivizes PMs to get through an improved milestone—often with a program that cannot be executed" (p. 301).

Mortlock and Dew (2021) describe dispositional optimism as the "tendency to expect positive outcomes even when such expectations are not rationally justified" (p. 98, citing Hmieleski & Baron, 2009). They note that "optimists expect good things to happen to them; they believe that chance events will break in their favor" (p. 98). Researchers have studied entrepreneurs because entrepreneurs deal with issues of uncertainty and long-time frames similar to those PMs face. These studies found that entrepreneurs are significantly more optimistic than non-entrepreneurs (Mortlock & Dew, 2021). Similarly, the DoD may select PMs from the pool of people willing to run large acquisition programs because those people feel too optimistic about their chance of failure (Mortlock & Dew, 2021).

Recency bias occurs when people give too much attention to the information they received most recently when making judgements. This behavior can look like overcorrecting for recent program overruns or successes, adopting an acquisition strategy because it is the latest focus of acquisition reforms rather than the best practice that has stayed relatively constant, or paying too much attention to new technology and not enough attention to historical trends of how long technology takes to become mature (Mortlock & Dew, 2021). Mortlock and Dew (2021) found evidence that recency bias was at play when they studied the Optionally Manned Fighting Vehicle program. The program's acquisition strategy implemented the newly introduced MTA pathway in a way that mirrored other programs that had failed (Mortlock & Dew, 2021).

Examining GAO reports on acquisition programs, Kiesling and Chong (2020) found evidence that supported many of the theories that cognitive biases had affected the decision-making processes and oversight assessments of these programs. While this study does not show causality between cognitive biases and program failures, it shows that the language surrounding these programs is consistent with what would be expected if cognitive biases were affecting their processes.

C. ORGANIZATIONAL AND CULTURAL FACTORS AMPLIFYING BIAS

Biases are present within complex environments comprised of organizational frameworks and cultural standards which together with incentive mechanisms can either support or suppress their influence. Mortlock and Dew (2021) identify the root causes of program failure as "ill-defined requirements, immature technology, integration challenges, poor cost estimating, unstable budgets, poor schedule planning, and schedule pressure from annual appropriation limitations" but argue that "an underappreciated reason for acquisition program failures and understudied part of big 'A' is the 'people part' of defense acquisition, which may have the largest effect on improving acquisition outcomes" (p. 96).

For example, Mortlock and Dew (2021) provide a framework for how cognitive biases are compounded by culture and leadership across three levels of the DoD decision-making apparatus. At the highest level, decisions occur at the institutional level and often follow a political model of decision-making in which bargaining between stakeholders with different preferences produces the outcome. One level down, at the Service or Program Executive Office level, decisions may depend more on whether an action appears to fit or go against the cultural norms of that organization. Program-level decisions most closely follow the rational actor model, with PMs considering pros and cons to choose the alternative with the best chance of success (Mortlock & Dew, 2021). Therefore, biases can play out differently depending on who is in the chain of command deciding. Cultural norms at each level can suppress or amplify individual biases. In addition, because each level feeds into the next, this interaction complicates matters by obscuring sources of bias (Mortlock & Dew, 2021). Efforts by PMs to rationally consider tradeoffs may be for naught if politically or culturally biased decisions have already been made upstream (Mortlock & Dew, 2021).

In particular, planners discount the likelihood that they experience cost or schedule growth throughout the course of a program. One reason for this tendency is that many elements of what can lead to good management judgment are biased themselves. Perceptions of control are at the heart of organizational management culture: managers are trained to believe, and held accountable in ways that reinforce the belief, that what

they do matters a great deal and directly determines the outcomes they experience. Planning for programs is built upon the overoptimistic views of the ability to control process and result. As such, many managers have an unrealistic belief in their level of control of both processes and results (Mortlock & Dew, 2021). Mortlock and Dew (2021) describe how "the planning fallacy creates biased expectations that mask a control gap that will exact a price over the course of most programs" (p. 98). Management practices that seem intuitive, such as focusing on program specifics and holding people accountable through incentives, actually "tend to encourage people to focus more intently on their plans, which increases bias" (p. 98, citing Buehler et al., 1997).

Common sense management practices can exacerbate biases. Intuitively, it makes sense that one should focus on the details of their specific program when planning but doing so reinforces an "inside view" of the decision and likely increases bias. Holding people accountable through incentives is another widely used management tool that research shows lead people to focus more on their plans, which again exacerbates bias (Buehler et al., 1997). The planning fallacy creates optimism biases that management practices reinforce, widening the gap between perceived control and actual control (Buehler et al., 1997; Mortlock & Dew, 2021).

Organizational culture can intensify escalation of commitment problems rooted in the sunk-cost fallacy. Hammond et al. (2006) warn that organizational cultures with disproportionately severe consequences for failed decisions create perverse incentives for managers to perpetuate failing projects indefinitely rather than acknowledge failure, hoping that continued investment will eventually reverse poor outcomes. Program cancellation can have large impacts on careers and can become highly politicized. This tendency can lead individuals to escalate their commitment to a program by throwing additional funds at it, rather than admitting (either to themselves or to others) that earlier estimates were wrong or that the original approach needs to be restructured (Hammond et al., 2006). Escalation of commitment thus continues as individuals act rationally within the constraints of the career consequences they face.

PMs are responsible for meeting cost, schedule, and performance objectives but do not control how those objectives are set (DoD, 2021). The Services often decide

performance requirements, cost objectives, and schedule targets and lock them down at major milestone decision reviews (Mortlock & Dew, 2021). In practice, this process can mean that PMs work with an APB that reflects biased decisions from others upstream, while Service-level or MDA decisions constrain them further (Mortlock & Dew, 2021; Drezner & Krop, 1997; GAO, 2015).

The GAO (2015) found that "PMs are incentivized to develop acquisition strategies focused on program approval at the milestone review but not acquisition strategies that could later be executed and deliver capabilities" (as cited in Mortlock & Dew, 2021, p. 97). The defense acquisition system "often provides incentives for Services and PMs to promote *successful* acquisition strategies (defined as approved and leading to successful milestones) rather than *sound* acquisition strategies (defined as executed within cost, schedule, and performance constraints, and leading to fielding capability)" (Mortlock & Dew, 2021, p. 107). Piling on optimistic assumptions to meet budget and schedule targets now can lead to cost and schedule growth down the line (GAO, 2015). Flyvbjerg et al. (2009) found that forecasting errors in large projects stem from "delusions or honest mistakes; deceptions or strategic manipulation of information or processes; or bad luck" (p. 172). They posit that the first two lead to consistent underestimation of infrastructure project costs (Flyvbjerg et al., 2009). Therefore, it appears that the factors which drive this type of forecasting error bias are likely to occur within any other large-scale bureaucracy that undertakes large scale planning and execution on a system with a significant amount of risk and public oversight. Under the deception model specifically, Flyvbjerg et al. (2009) found that "politicians, planners, or project champions deliberately and strategically overestimate benefits and underestimate costs" (p. 173) in order to increase the likelihood that their projects gain approval and funding.

D. AI AND MACHINE-ASSISTED DECISION SUPPORT

The previous sections established that cognitive biases systematically distort acquisition decisions and that organizational culture and incentive structures amplify rather than correct those distortions. The following sections examine whether artificial intelligence and machine-assisted decision support offer a remedy by reviewing AI's

relevant technical foundations, its potential advantages for acquisition planning, and the risks and limitations that accompany its use.

1. AI Capabilities Relevant to Decision-Making

AI can be generally described as the science and engineering of making intelligent machines and computer programs that are able to complete tasks requiring intelligence if done by humans (Koszykowski & Orzeszko, 2025). ML is a subset of AI focused on algorithms that identify patterns in data and use what they learn to complete tasks rather than follow static rules (Koszykowski & Orzeszko, 2025; Narbaev et al., 2024). Common ML algorithms include "neural networks, decision trees, support vector machines, and ensemble methods" (Mini, 2026, p. 1), which researchers have applied across various project scheduling subprocesses (Koszykowski & Orzeszko, 2025). Generative AI refers to more recent advancements in ML, such as large language models or LLMs that can generate text, assess moves in games, and create new outputs that it did not directly learn during training, but based on patterns and relationships it has learned (Csaszar et al., 2024; Miedema et al., 2026). Csaszar et al. (2024) differentiate between weak AI, which is where current LLMs fall under and are able to achieve performance on par with humans for particular tasks, and strong AI, which would require developments in causal reasoning abilities and knowledge transferred between different domains that are not present (Csaszar et al., 2024; Koszykowski & Orzeszko, 2025). All AI discussed in this study fall under weak AI heading.

While AI and ML models operate on information very differently than the human mind does, it may allow them to make decisions with less bias. Harris (2020) notes that "machine learning algorithms were designed to make decisions not only faster but also more accurately and fairly" and that "these algorithms are designed to eliminate or reduce cognitive biases" (p. 1; see also Csaszar et al., 2024; Miedema et al., 2026). Because AI does not experience the information overload that humans do, it does not have to take the cognitive shortcuts that produce human bias (Csaszar et al., 2024). They will use the same decision logic from one case to the next, regardless of mood, tiredness, or other recently made decisions (Echterhoff et al., 2022). And they can identify patterns to alert

humans to relationships in past data they may not have noticed (Narbaev et al., 2024; Stebbins et al., 2024).

Defense organizations are already using AI to conduct some decision-support tasks (Stebbins et al., 2024). Research conducted by the RAND Corporation explored how "AI/ML tools and platforms may enable (or hinder) existing Army IPB processes" (Stebbins et al., 2024, p. 2). The study found that "AI has the potential to mitigate bias in human decision making by using algorithms that are intentionally designed to mitigate bias" and that machine-enabled teams could "foster creativity, challenge analytical assumptions, and promote a reflexive learning environment" (Stebbins et al., 2024, pp. 55, 59).

Outside of defense-related tasks, AI and ML have also been used in predictive maintenance within manufacturing as well as other decision-support operational uses across different industries (Ayvaz & Alpay, 2021; Miedema et al., 2026). Businesses across industries are also increasingly using AI tools to analyze and filter data for human decision-makers in areas like finance, healthcare, and supply chain optimization (Miedema et al., 2026). The fact that these systems are used to support decisions in complex, real-world environments shows that AI can match humans in some of the tasks necessary for addressing cognitive biases in national security contexts.

Research outside of the defense industry shows that researchers and practitioners have used AI to perform functions similar to, or proxies for, tasks involved in acquisition planning. Researchers have applied ML techniques such as neural networks, decision trees, and ensemble models to project cost estimation (Koszykowski & Orzeszko, 2025; Narbaev et al., 2024). Studies have found these tools can surpass traditional earned value management equations by learning from historical cost performance data across a variety of past projects to produce more accurate estimates (Narbaev et al., 2024, p. 4373). Another literature review analyzed the application of ML to three aspects of project schedule development: effort estimation, activity sequencing, and duration prediction (Koszykowski & Orzeszko, 2025). The review found there is evidence that properly trained ML algorithms can support all three of these tasks, and in some cases outperform rule-based methods (Koszykowski & Orzeszko, 2025). These studies have largely

occurred in the fields of construction and software engineering (Koszykowski & Orzeszko, 2025; Narbaev et al., 2024). However, the tasks of cost estimation, schedule forecasting, and risk analysis under uncertainty are directly analogous to and are integral parts of the defense acquisition planning process.

Researchers at the RAND Corporation conducted another study with potential application to countering bias in acquisition planning (Stebbins et al., 2024). The study focused on a subset of intelligence analysis specifically related to IPB processes in the U.S. Army (Stebbins et al., 2024). Analysts who perform IPB research about adversaries and generate intelligence reports can be subject to many of the same biases described earlier in this paper. The researchers sought to understand whether an AI-augmented process could counter potential biases by processing vast amounts of data to identify patterns, alternative courses of action, and other insights that a human analyst might miss (Stebbins et al., 2024). Ultimately, the study determined that AI applications could support human decision making by surfacing gaps in analyst knowledge (Stebbins et al., 2024). Stebbins et al. (2024) found that AI platforms served as a "useful 'tipping' mechanism to help IPB stakeholders target specific information gaps and sources" (p. 57). The machine-enabled team was able to surface "additional civilian groups, local-level NGOs, and stories beyond adversaries" that the human-only team did not identify, demonstrating "a nascent ability of AI to foster double-loop learning within the IPB process, in turn allowing for a reflexive approach to mitigating potential human bias" (p. 62).

Studies on how humans interact with computers have identified AI methods that are used to recognize and eliminate cognitive biases from decision-making. Echterhoff et al. (2022) studied the manifestation of the Anchoring Bias when people are asked to make a series of related decisions. They found that the anchoring bias does not stop at one single judgment; it occurs repeatedly across subsequent judgments causing the evaluator to develop an implicit bias based upon their previous evaluations (Echterhoff et al., 2022). To determine whether an AI system could avoid this bias, the researchers trained a support vector machine (SVM) on college admissions and product review data, reasoning that they required "a method not subject to anchoring bias" because ML algorithms "do not have access to the specific ordering of the files" (Echterhoff et al.,

2022, p. 3). Because the SVM algorithm did not have access to the sequence of past decisions, anchoring bias did not affect it (Echterhoff et al., 2022). The researchers then modeled the probabilistic state of an evaluator's anchoring bias and used reinforcement learning to test which sequences of decisions were optimal for presenting to the SVM (Echterhoff et al., 2022). By learning the optimal order in which to present instances for review, the researchers achieved "an increased agreement to ground-truth by 7% and reduced bias of 0.07" (Echterhoff et al., 2022, p. 8), while the retrospective probabilistic adaptation strategy separately increased accuracy by 2–5% (Echterhoff et al., 2022, p. 2). This research has clear applications for acquisition decision-making, as the process requires PMs to make a series of judgments about requirements, alternatives, risks, and costs. Like the example decision sequences used by Echterhoff et al. (2022), each decision could theoretically be used to anchor subsequent judgments. The research shows that AI can potentially improve decision-making environments (Echterhoff et al., 2022).

Research in fields outside of the military and defense shows that AI systems can either assist or reach human-level performance in certain strategic decision-making tasks (Csaszar et al., 2024). In one study, Csaszar et al. (2024) analyzed data from a top-tier startup accelerator as well as a business plan competition to see how entrepreneurs and investors performed on tasks related to business planning when compared with LLMs. The researchers found that LLMs were able to create and assess business strategies with performance like that of human entrepreneurs and investors (Csaszar et al., 2024). Csaszar et al. (2024) found that in a "weak AI world" (the current state of LLMs), AI can "help decision makers quickly generate and evaluate more strategic alternatives than they could otherwise consider, leading to better average decisions, especially among lower-performing experts" (p. 333). However, AI will not "magically produce new strategies that would be inaccessible to high performing humans using conventional analysis methods" (p. 333). This conclusion has important implications for acquisition planning. If the automatic analysis of plans can improve the performance of good decision-makers but only moderate the performance of bad ones, it stands to reason that AI has the greatest potential to improve decision quality not by replacing human judgment but by improving the decisions of those worst impacted by biases.

There is one final body of literature that frames this discussion by considering the different roles AI systems can play in supporting or completing decision-making tasks. Miedema et al. (2026) categorize AI systems into three distinct types, based on how much they can make decisions. These categories are autonomous systems, which operate independently of people; supportive systems, which help human users by providing decisions and suggestions; and collaborative systems, which work with human teams to accomplish goals (Miedema et al., 2026, p. 1).

Considering this framework, it becomes clear that AI will not replace human acquisition planners, at least for the near future. Given the stakes of acquisition decisions, including legal requirements, financial consequences on the order of billions of dollars, and implications for national security, some degree of human judgment is necessary. As a result, AI will likely be used to complete one of the other two roles described by Miedema et al. (2026). It is worth noting that these roles are not mutually exclusive and AI could complete all three functions in different portions of the acquisition process. However, for the purposes of this research, it is useful to think of AI as either generating recommendations for human judgment (supportive) or evaluating decisions made by humans (collaborative).

2. Potential Advantages of AI in Acquisition Planning

There are several features of AI systems that point toward potential advantages for use within defense acquisitions. First, AI systems differ fundamentally from humans because they do not experience cognitive biases like the ones previously discussed (Csaszar et al., 2024; Miedema et al., 2026). AI systems don't see increased perceptions of control when engaging in detailed planning (the planning fallacy) (Csaszar et al., 2024; Miedema et al., 2026). AI systems don't overweigh important but uncommon recent experiences (availability bias) (Csaszar et al., 2024; Miedema et al., 2026). AI systems aren't attached to an initial number estimate the way humans are (Echterhoff et al., 2022; Tversky & Kahneman, 1974). AI systems can have biases and what they produce reflects that bias in their training data or model architecture (Miedema et al., 2026; Stebbins et al., 2024). Unlike human cognitive bias, which operates unconsciously and "cannot be turned off at will" (Kahneman, 2011, p. 25), AI developers can identify AI bias through

testing and mitigate it through data curation, model validation, and human oversight (Miedema et al., 2026; Stebbins et al., 2024).

Second, AI systems reason systematically within the structure they receive (Csaszar et al., 2024; Miedema et al., 2026). For example, recent personally relevant failures, organizational pressures to reduce estimated costs and timelines, and innate optimism or pessimism may all impact one PM's judgment. Two PMs' judgments should therefore vary to some degree (Kahneman & Lovallo, 1993; Mortlock & Dew, 2021). An AI system trained with the same model applies consistent weighting to small-probability high-impact risks the same way for each program, resulting in less variance in quality from one program to the next (Csaszar et al., 2024; Miedema et al., 2026). This consistency could prove particularly useful for baseline development efforts if developers can use historical data on comparable programs to train AI systems to produce more realistic estimates (Narbaev et al., 2024). Research on ML models for project cost estimation suggests that training AI on historical program data to improve cost forecasting accuracy is possible (Narbaev et al., 2024). Narbaev and colleagues (2024) found that ML algorithms trained on past project performance data to generate cost forecasts outperformed baseline linear estimation formulas; one reason cited for the superiority of ML methods was their ability to "capture nonlinear patterns in cost performance . . . as well as complex interactions between variables that traditional methods may overlook" (Narbaev et al., 2024, p. 4385). Adapting this general approach for use in defense acquisition baseline development could guard against the kind of systematic optimism that the GAO has documented across programs for years, but needs to be designed carefully (GAO, 2020; Narbaev et al., 2024).

Third, AI systems have the ability to process more information than a human could realistically review (Csaszar et al., 2024; Miedema et al., 2026). Therefore, at scale, AI can find correlations among historical data from programs and connect the key elements of a program to previous programs that had comparable outcomes and trends and locate and highlight high-risk areas which may not be visible by reviewing documents manually (Narbaev et al., 2024; Stebbins et al., 2024). In addition, when applied to the risk assessment and alternative evaluation processes, this capacity will allow for the review and consideration of many more alternatives and risks than are

currently reviewed by humans through a manual process (Csaszar et al., 2024). Csaszar and colleagues (2024) describe this capability as AI's potential to extend the limits of bounded rationality. When integrated into the decision-making process, AI can "augment the fundamental cognitive processes of search, representation, and aggregation that support human strategic thinking," allowing humans to create and evaluate strategies more quickly, thoroughly, and at a larger scale than would otherwise be possible (Csaszar et al., 2024, p. 322). Given the same case study on which to base a program's business case, an AI system could potentially identify more strategies than a human finds reasonable, surface risks that a single PM doesn't personally encounter across their career, or identify historical inconsistencies between the assumptions driving program planners' estimates and the actual performance of similar programs in the past (Csaszar et al., 2024; Narbaev et al., 2024).

Human-AI interaction research also hints at the potential for AI to mitigate human bias. Harris (2020) notes that "machine learning algorithms were designed to make decisions not only faster but also more accurately and fairly" (p. 1), while Echterhoff et al. (2022) demonstrated that AI can detect and correct bias in sequential human decisions. Research exploring how AI might combat anchoring bias finds that ML models, which do not see the intended decision's placement or position among other decisions, in other words, do not have "access to the ordering information which may be anchoring human reviewers" (Echterhoff et al., 2022, p. 3) and hence are not subject to the sequential dependencies that cause anchoring. Early research using AI to mitigate cognitive bias has taken two broad approaches relevant to acquisition planning. The first focuses on debiasing the algorithm itself: Harris (2020) found that combining pre- and post-processing fairness algorithms "mitigat [ed] biases effectively, and these were most effective when they are combined" (p. 8). The second focuses on debiasing human decisions by using AI to detect and correct bias in real time (Echterhoff et al., 2022). Harris (2020) demonstrated a complementary approach: when researchers trained a machine learning algorithm on biased human ratings, it "propagat [ed] the biases" (p. 1), but applying algorithmic fairness techniques at the various processing stages could recover nearly all lost accuracy. The best combination of debiasing algorithms achieved a prediction accuracy of 0.949, approaching the unbiased baseline of 0.952 (Harris, 2020,

p. 8). This suggests that developers could remove known biases in AI systems used in acquisition planning before their outputs reach human decision makers.

Echterhoff et al. (2022), on the other hand, showed that researchers could also use AI to structure the decision-making process in a way that prevents bias before humans make decisions. The researchers proposed that AI could either retrospectively adapt decisions when it detects anchoring or prospectively reorder the sequence to prevent bias. In the retrospective case, they found they could "mitigate bias retrospectively for already-made decisions by capturing the anchoring state of a reviewer" (Echterhoff et al., 2022, p. 8), a model the authors noted could also be used to flag potentially biased instances for re-review (Echterhoff et al., 2022, p. 7). In the prospective case, they used reinforcement learning to present instances in an order that minimizes anchoring effects before the reviewer makes the decision (Echterhoff et al., 2022, p. 4). In an acquisition context, a similar system could track the order in which a program office reviews a list of cost elements or alternative approaches to acquiring a capability (Echterhoff et al., 2022; Mortlock & Dew, 2021). When the pattern of judgments suggests they may be anchored to an initial estimate or influenced by the cost or viability estimate entered immediately before it, the system would flag the current estimate under review for reevaluation (Echterhoff et al., 2022). This type of AI application would not replace human judgment; it would instead use human judgment where it is highest and apply AI's pattern-recognition capabilities to identify decision-making patterns that individuals are unable to see in their own behavior (Echterhoff et al., 2022; Miedema et al., 2026).

Csaszar et al. (2024) offer a helpful framework for thinking about AI's potential to improve decision-making by distinguishing between the implications of weak AI versus strong AI. In a weak "AI world" (the world we live in with LLMs), they argue AI can "help decision makers quickly generate and evaluate more strategic alternatives than they could otherwise consider, leading to better average decisions, especially among lower-performing experts" (p. 333). But AI will not "magically produce new strategies that would be inaccessible to high performing humans using conventional analysis methods" (Csaszar et al., 2024, p. 333). Strong AI, on the other hand, which would require making breakthroughs in AI's capacity for causal reasoning and applying knowledge from one domain to problems in another, could conceivably identify options

that would be impossible for human managers to realize (Csaszar et al., 2024). For acquisition, most of the research discussed here is happening under the weak AI paradigm. The question is not whether AI can outperform a seasoned PM at identifying risks or writing insightful program histories but whether AI can help overcome biases that impact all PMs to some degree and improve the average quality of plans.

3. Challenges and Risks of AI Use

AI is not without risks of its own. Training data, model topology, and even prompt construction can bias AI systems (Miedema et al., 2026). An AI system specifically trained on historic acquisition decisions may simply replicate prior biased decisions instead of providing independent judgment (Stebbins et al., 2024). Moreover, developers must take care to curate AI training data properly and to test generated output for bias (Miedema et al., 2026; Stebbins et al., 2024). This bias limitation of data-driven AI, noted by Miedema et al. (2026), arises because data-driven AI systems depend on "large volumes of high-quality data" yet "in practice, such datasets are often limited, expensive, or difficult to obtain" (p. 7). This characterization of datasets aligns with the fragmented state of past defense acquisition data, which exists in silos by program, security classification, and organizational database (Miedema et al., 2026). Additionally, "data-driven AI does not inherently incorporate domain knowledge," which "increases the risk that the generated outputs deviate from established industry standards, best practices, or constraints" (Miedema et al., 2026, p. 7). Therefore, there may be issues with outputs that stray from established policy, statutory guardrails, or what other experienced professionals would understand to be non-negotiable.

Miedema et al. (2026) "distinguish between transparency, interpretability, and explainability," defining "transparency as the extent to which the inner workings of a model are open and inspectable," interpretability as "how easily a human can comprehend the model's behaviour or logic," and explainability as methods "developed to provide insight into AI systems that are neither transparent nor easily interpretable" (p. 7). Many modern AI systems, particularly LLMs that rely on deep learning, generate outputs that appear correct but that humans cannot explain in terms of how the model arrived at them (Miedema et al., 2026). In many acquisition decisions, program managers

must provide justifications to a sponsoring organization, Congress, or some other body (DoD, 2021, 2022b). These stakeholders may not accept or use decisions that lack transparent justification and may not accept AI-generated justifications that they cannot understand. Stakeholders involved in the acquisition planning process, including MDAs, GAO investigators, and congressional committees, require transparency, interpretability, and explainability before accepting decisions supported by AI.

Miedema et al. (2026) report that "explanations provided by AI increase the likelihood of acceptance of AI's recommendations by humans, regardless of whether the recommendation is correct" (p. 8). Essentially, some explanation was better than none when it came to persuading people to accept decisions. Trust and human adoption are also challenges to the responsible use of AI (Miedema et al., 2026). Studies on automation bias have found that humans using AI can over- or under-trust the systems they are using (Alon-Barkat & Busuioc, 2023). To better understand automation bias in public decision-making contexts, Alon-Barkat and Busuioc (2023) ran three survey experiments with an aggregated sample of 2,854 participants. Their "experimental findings from three separate studies…do not reveal a general pattern of automatic adherence to algorithmic advice" (Alon-Barkat & Busuioc, 2023, p. 164). The authors attributed this absence of automation bias in part to "a relative skepticism about the performative capacity of AI algorithms" among participants who remained under-exposed to algorithmic systems in practice (Alon-Barkat & Busuioc, 2023, p. 165). These researchers did find instances of what they describe as selective adherence. In their words, decision-makers were "likely to rely on algorithmic inputs in a biased, selective manner—to assign more weight to the advice…when this is aligned with pre-existing stereotypes" (Alon-Barkat & Busuioc, 2023, p. 165). Applied to acquisition, PMs may follow AI recommendations that agree with established courses of action and ignore AI recommendations that provide alternatives to cognitive biases (Alon-Barkat & Busuioc, 2023). For example, a PM may have a pre-existing preference for a certain contracting strategy. If given an AI decision that supports that strategy, the human decision-maker may accept it without question. However, if the AI suggests a different contracting approach, the PM may disregard the AI output, instead doubling down on their original

decision. To avoid this tendency, human users must neither distrust nor completely trust AI.

During one of Alon-Barkat and Busuioc's (2023) studies, the Netherlands childcare benefits scandal took place. The scandal involved the Dutch tax authorities' reliance on "a 'learning algorithm' that used, among other criteria, nationality as a discriminant predictive feature" to flag high-risk applicants (Alon-Barkat & Busuioc, 2023, p. 162). Furthermore, as the authors noted, "both the automated risk selection and the individual investigations of officials were discriminatory" (Alon-Barkat & Busuioc, 2023, p. 163, citing Volkskrant, 2020). As Alon-Barkat and Busuioc (2023) concluded, "the scandal speaks acutely to the serious real-life repercussions that can arise when human bias meets algorithmic bias in bureaucratic decision making" (p. 166). The authors warned that "bias can also crop up at another level: in the human–AI interaction, in how decision-makers process, interpret, and act upon algorithmic outputs" (Alon-Barkat & Busuioc, 2023, p. 166). If historical acquisition data influences an AI system's decisions, there is just as easy a pathway for algorithmic discrimination to manifest in program decisions. If the historical data showed that most programs used cost-plus contracts, the AI may favor that contracting method in its decision. If the PM holds a bias toward cost-plus contracting, they may follow the AI's decision. However, it is equally possible that the AI decision goes against the PM's bias, who may then ignore the AI decision. To help account for these possibilities, this study compares decisions across three separate AI models and utilizes an evaluation rubric that flags potential signs of cognitive bias.

It is also important to note that AI systems hallucinate (Miedema et al., 2026). In acquisition decisions that may cost billions of dollars and impact national security, decision makers cannot accept AI hallucinations at face value. Miedema et al. (2026) note that "LLMs can produce factually incorrect outputs and are prone to so-called hallucinations" (p. 9), a risk amplified in specialized domains where "missing or outdated knowledge in training data can lead to wrong recommendations" (p. 17). If there is a gap in the training data or the knowledge domain of an AI, there is the potential for AI to make suggestions that are against policy, violate statutes, or are otherwise impossible

given the particulars of a program (Miedema et al., 2026). Organizations need safeguards in place to review AI decisions for accuracy.

In many ways, accountability for decisions presents a separate challenge than AI and human interaction. Miedema et al. (2026) identify accountability, transparency, and controllability among the core "dimensions of trustworthiness" for AI systems (p. 6), emphasizing that "controllability requires that the use of AI and its outcomes remain under human agency when needed" (p. 16). The acquisition process holds people accountable. Decision authorities are responsible for the programs they choose to approve, PMs are responsible for costs, schedules, and performance, and statute requires program offices to receive an independent cost estimate to check their estimates (DoD, 2021; Weapon Systems Acquisition Reform Act of 2009, Pub. L. No. 111–23). Responsibility for poor decision-making by AI systems remains ambiguous because it is not clear who should be held accountable when these systems make mistakes (Miedema et al., 2026). Is it the PM who accepted it? The organization who chose to use the AI for decision-making? The developers who trained the model? Until policy clearly answers these questions, acquisition professionals have a reason to avoid AI that could shift or eliminate responsibility.

E. GAPS IN EXISTING RESEARCH

The present study seeks to fill several gaps in research literature. First, there is no known body of work focused on acquisition decisions made by or with the assistance of algorithms. While the DoD has applied AI and ML technology in multiple defense decision-making contexts, algorithmic decision-support research in the academic literature has focused primarily on commercial goods procurement or public sector decision-making more generally rather than acquisition constrained by laws, regulations, and policies specific to the DoD. Many features unique to defense acquisition create a decision space unlike that of most commercial or public sector work. These include the statutory requirements, stakeholder equities, risk tolerance, and industrial base considerations that constrain options. Similar research, including ML for project cost estimates (Narbaev et al., 2024) and schedule development (Koszykowski & Orzeszko, 2025), has largely occurred in construction and software engineering contexts, has not yet

adapted to defense acquisition. These require validation against DoD data because of differences in scale, oversight, regulation, and policies to ensure they meet requirements.

Second, although researchers have observed cognitive bias in organizational decision-making in general, documented associations between cognitive biases and the decisions that lead to the formation of ASs and APBs are limited. Evidence of cognitive bias in ASs or APBs appears to be purely circumstantial and based on observations of programs rather than empirical measurement of bias in plans. Mortlock and Dew (2021) "studied three defense acquisition programs and found strong evidence that systemic behavioral biases affected the management and decision-making within these programs" (p. 111), their conclusions rested on retrospective case studies of program execution history rather than controlled comparisons of biased plans to unbiased ones. Prior to that behavioral bias analysis, Mortlock (2020) had demonstrated through a prospective survey design that acquisition professionals formulate strategy components in highly variable ways when given identical programmatic inputs. In that study, Mortlock provided 31 professionals with the same JCM Milestone B data, requirements, technology readiness levels, risk ratings, and competing cost estimates, and asked them to recommend an acquisition strategy. The results produced a wide variety of recommended strategies, with none of the 31 respondents recommending an approach resembling the incremental strategy ultimately adopted by the Services over a decade later (Mortlock, 2020). This finding suggested that the variability itself may be a product of cognitive biases rather than legitimate analytical differences, motivating the behavioral acquisition research agenda. However, neither Mortlock (2020) nor Mortlock and Dew (2021) introduced a non-human comparison group against which to measure whether the observed variability and bias indicators were unique to human cognition or inherent to the decision environment itself. This study fills this gap by using AI as that comparison group, administering the same survey instrument that was previously given to human professionals, to AI models to isolate the human variable.

Third, no one has yet described a unified methodology for introducing AI into acquisition planning. Although tools exist and experiments are underway within defense organizations, no one has articulated a process for how program offices would create, validate, and introduce AI recommendations into the formal acquisition decision-making

process. Is an AI recommendation simply another opinion that decision-makers weigh alongside subject matter experts? Who is accountable for the decisions made by or with the assistance of an AI tool? How do we operationalize "responsibility and authority" for decisions supported by algorithms within the existing acquisition regulations? Broader work around human interaction with AI systems in public sector decision-making has started to tackle some of these questions. Alon-Barkat and Busuioc (2023) study how public sector decision-makers interpret and use algorithmic advice. Miedema et al. (2026) propose a life cycle framework for trustworthy AI decision-making. Yet neither of these studies has focused on, nor has anyone adapted either to, environments with structures like those governing defense acquisition programs, where a formal hierarchy of MDAs reviews each plan against an independently generated cost estimate before approving programs to be reported to Congress.

Fourth, no known studies have empirically compared human and AI acquisition decisions made using the same inputs. It is unclear whether acquisition decisions that AI makes or helps make will be more or less accurate, objective, or consistent than decisions humans make alone. We don't know if AI introduces the same biases as humans or different/better/more problematic ones. We don't know what kinds of strategies and baselines an AI would choose given the same information that a human PM would have. Csaszar et al. (2024) made strides in this area by generating and comparing human and AI-created strategic decisions in the business planning context, but to my knowledge, no study has offered this kind of fundamentally similar comparison in a defense acquisition context. Without such a comparison, we cannot know AI's value for acquisition planning.

Finally, while there is research showing that AI can identify and mitigate cognitive bias in laboratory and crowdsourced settings (Echterhoff et al., 2022) and improve strategic decision quality relative to humans alone (Csaszar et al., 2024), there has been no study of whether these findings hold true in defense acquisition. Work environments, incentive structures, and consequences for errors vary greatly between the examples where studies have shown AI to mitigate cognitive bias (e.g., college admissions review processes, product evaluation, startup strategy development) and defense acquisition planning. Will algorithmic techniques proven to mitigate anchoring effects on university admissions committees also mitigate anchoring in a program office

working through trade-offs to establish a cost baseline that represents billions of dollars of future taxpayer obligation? How will AI tools contend with multiple layers of organizational cognitive bias when developing or reviewing acquisition plans? To what extent are the biases affecting algorithmic recommendations different from those affecting the human decision makers that design and use them? By controlling for information available to human decision makers and requiring AI and human participants to complete the same decision tasks, this research answers those questions.

F. SUMMARY

Taken together, the research reviewed clearly and compellingly establishes both the problem and the need for this research. Cognitive psychology research spanning several decades has clearly shown that human judgment follows predictable patterns of bias, and that even experienced professionals who have relevant training and who work in environments with high stakes and access to data and decision support tools remain susceptible to these biases (Gilovich et al., 2002; Kahneman, 2011; Tversky & Kahneman, 1974). The defense acquisition workforce is by no means exempt from these findings (Kiesling & Chong, 2020; Mortlock & Dew, 2021). Mortlock and Dew (2021) concluded that the prevalence and persistence of the four biases across the programs they studied highlights which specific cognitive distortions are most significant in defense acquisitions. Organizational research has shown that "the culture and leadership at different levels of the DoD from the institutional level to the organizational level to the program level affect the impact of the biases" (Mortlock & Dew, 2021, p. 111). When these factors combine, they create an acquisition environment where biased estimates and decisions become the norm rather than the exception (Mortlock & Dew, 2021).

The research reviewed also provides reason to be hopeful that AI can play a role in addressing these challenges. AI/ML is not subject to the same cognitive limitations as humans. These systems can learn from far more historical data than any person could reasonably process and has already shown success improving decision-making outcomes in related fields such as cost estimation, scheduling, and planning (Csaszar et al., 2024; Koszykowski & Orzeszko, 2025; Narbaev et al., 2024). In fact, one study investigating the ability of ML to detect and correct anchoring bias in sequential decision-making tasks

showed promise in an application with direct relevance to acquisition planning (Echterhoff et al., 2022).

This is not to say that the application of AI will be without its own hurdles. If developers train AI systems on acquisition data with embedded biases, those systems risk repeating those mistakes (Miedema et al., 2026; Stebbins et al., 2024). Concerns about AI explainability, hallucinations, unclear lines of accountability, and selective compliance pose additional challenges (Alon-Barkat & Busuioc, 2023; Miedema et al., 2026). And of course, while studies have shown AI to be effective in mitigating bias in other contexts, none have occurred in the specific context of defense acquisition, which features unique decision-making structures, statutory requirements, and financial stakes in billions of dollars.

Filling these gaps in the current research landscape is the core objective of this research. Requiring humans and AI to perform the same acquisition planning tasks given the same information allows this research to make the first direct comparison of human and AI judgment in acquisition and determine whether AI can reduce cognitive bias that has plagued the defense acquisition process and contributed to cost growth, schedule slip, and performance shortfalls.

IV. RESEARCH METHODOLOGY, DATA COLLECTION, AND ANALYSIS

This chapter reports on the execution of the comparative case study research introduced in Chapter I. The research compares human-produced acquisition planning outputs against AI-generated outputs. This is accomplished using the same case study data, constraints, and survey instrument to systematically identify differences in decision quality, analytical rigor, evidence of cognitive bias, and internal consistency. Chapter VI walks the reader through the complete research process, including the case study and survey instrument, human and AI data collection procedures, the model selection rationale and prompt design, the evaluation rubric, the comparison framework and statistical methods, and finally the results of the analysis. This chapter reports and compares findings on decision quality, evidence of cognitive bias, and analytical characteristics between human acquisition professionals and multiple AI systems.

A. THE CASE STUDY: JOINT COMMON MISSILE PROGRAM

The primary research instrument is a substantive acquisition case study centered on the Joint Common Missile (JCM) program. The JCM is an actual Defense Acquisition Program classified as an ACAT-1D. It also provides copious historical data to draw from and has significant levels of observed complexity that are well documented throughout its history. This research selected the case because it represents a system where cognitive biases known to occur during acquisition planning should be present, given the level of technical complexity, number of stakeholders with competing interests, existing cost and schedule tension between internal program estimates and external independent estimates, and integration risk spanning multiple platforms and services. Research has shown that human decision makers fall back on heuristics most frequently when dealing with high levels of complexity and uncertainty (Tversky & Kahneman, 1974) and multiple recent studies have shown that conditions which produce high levels of complexity and uncertainty produce observable planning fallacy, over-optimism, trade-off difficulty, and recency bias effects in real-world acquisition decision-making (Mortlock & Dew, 2021).

This study selected the JCM case specifically because Mortlock and Dew (2021) found through analysis that the planning fallacy, difficulty making trade-offs, and over-optimism bias influenced the JCM's decision-making. Mortlock (2020) first used the JCM program as a case study framework for acquisition strategy research, surveying 31 acquisition professionals with actual JCM Milestone B decision data. The survey, which Mortlock developed based on the work of Gress, Kohtz, and Noll (2018) at the Naval Postgraduate School, provided participants with the actual programmatic data that the PM used, program management office, program executive offices, Service Acquisition Executives, and Milestone Decision Authority. The survey organized inputs into three categories: technology, requirements, and resources. Participants then recommended whether to pursue a single-step, two-increment, or three-increment development approach (Mortlock, 2020). This study extends Mortlock's (2020) research by administering the same survey to AI models and comparing the AI-generated strategy recommendations against the human baseline data he collected.

The case study presents the details of the JCM program at a snapshot in time. The program is roughly six months out from the planned Milestone B (initiate EMD) decision. The planned EMD phase lasts four years and targets initial operational capability (IOC) five years out from Milestone B. The program is a joint service program between the Army and Navy and is meant to field a replacement weapon system for four separate platforms: AH-64 Apache, AH-1Z Cobra, MH-60 Seahawk, and F/A-18 E/F Super Hornet. The approved CDD contains KPPs that include: a tri-mode seeker (precision point, active, and passive), a multi-purpose warhead, and common motor. All technologies listed as critical technology elements (CTE) sit at Technology Readiness Level (TRL) 6, having matured through Science and Technology Objectives (STO) and Technology Maturation and Risk Reduction (TMRR) efforts.

This study selected the scenario because it contains engineered tension points intended to produce competing sources of data to create conditions that would allow anchoring to be observed in survey responses. These tensions were selected because they are real tensions that were present on this program and exist broadly across ACAT-1 programs. Schedule and cost present the greatest tensions. The draft APB calls for a 48-month EMD schedule and $108-120K AUPC. The CAIG ICE indicates a possible

schedule range of 72–144 months for EMD and $153K AUPC. These differences are not subtle. The program's internal estimate is 33–67% shorter on schedule and 21–29% lower on cost than the independent estimate. When two authoritative data points exist in tension, internally generated and optimistic, externally imposed and conservative, respondents will necessarily have to anchor to one data source or the other. The extent to which they allow tension to influence their selections is quantitatively observable optimism bias.

This fundamental tension also exists between the scholarly literature and historical trends across defense acquisition. The GAO has consistently identified overly optimistic cost and schedule baselines as the single greatest driver of program underperformance (GAO, 2015, 2020, 2025), while Flyvbjerg et al. (2009) demonstrate both cognitive biases and evidence of intentional misrepresentation led to systematic underestimation of project costs on large programs. Specifically, the GAO Cost Estimating and Assessment Guide details how ICEs are intended to exist as an independent and objective measure to provide realistic assessment of program cost achievability (GAO, 2020). As such, the existence of the CAIG ICE in this scenario is analytically intentional. The risk assessment explicitly supports the presence of these tensions. Integrated system risk at Milestone B is projected to be Medium-High as a function of five risk items relating to propulsion integration, seeker integration, lethality of warhead, common motor across multiple air platforms, and software memory requirements. Despite this risk, the program office plan calls for the longest possible schedule supported by any source of truth (the APB at 48 months). A respondent who selects the shortest schedule option while also indicating that risk assessment applies to this decision has demonstrated a logically inconsistent pattern of selections that directly relates to theories of the planning fallacy described in Section 2.

According to Mortlock and Dew (2021), the planning fallacy arises when the planning process itself biases the beliefs of managers responsible for producing program estimates, causing them to make decisions that are too optimistic about future performance. Kahneman and Lovallo (1993) explained this mechanism: forecasters adopt an inside view that "is generated by focusing on the case at hand, by considering the plan and the obstacles to its completion, by constructing scenarios of future progress, and by

extrapolating current trends" (p. 25). By developing such a detailed plan, managers' perceptions of control are enhanced, leading to unwarranted confidence in their forecasts (Mortlock & Dew, 2021, p. 97). The GAO (2015) even found PMs were explicitly incentivized to produce acquisition strategies that ensure programs advance past milestone reviews, not ones that could be successfully executed and deliver capabilities. These conditions programmatically incentivize the behavior described by the Kahneman and Lovallo model.

The response selections around performance parameters add additional depth. As written, the CDD KPPs demand the tri-mode seeker, multi-purpose warhead, and common motor. However, survey respondents can technically choose to build a lower performing missile by only selecting single-mode or dual-mode seekers, requiring only one COTS warhead from an existing missile family (Hellfire, TOW, or Maverick), or using a single COTS motor. Each of these selections represent trades that could hypothetically be made by a program manager to lessen perceived risk, cost, or schedule. If a respondent chooses to select the KPP or deviate from it, and how they answer follow-on questions about the program's reaction to KPP relief, will show how strongly fidelity to requirements inhibits risk-taking and trade-off behaviors. According to Mortlock and Dew (2021), inability to make proper trades is one of four primary biases present in decision making across defense acquisition programs. They explain that when programs have multiple requirements that are framed in such a way that they cannot be logically compared to one another, decision makers feel the need to satisfy all requirements rather than deciding which ones are most critical. JCM was one of these programs; rather than deciding between which Service or air platform should receive capability first, the program tried to equip everyone at once (Mortlock & Dew, 2021).

B. DATA COLLECTION

Now with the case study information established, the next step involved collecting comparable decision outputs from both human acquisition professionals and AI systems using that same data. The following sections detail how each data set was gathered, the controls applied to ensure valid comparison, and the evaluation framework used to score both sets of outputs

1. Human Data Collection

The study drew human response data from defense acquisition professionals who previously completed the JCM Acquisition Strategy Survey (Mortlock, 2020). Mortlock (2020) designed the survey so that each participant had enough data to select survey responses that build upon each other to form an appropriate acquisition strategy. The survey organized inputs into three categories: technology (WBS risk ratings and CTE TRLs), requirements (CDD KPPs, IOC, AO, and AUPC presented in a draft APB), and resources (POE and ICE cost estimates). Participants then choose whether to recommend a single-step, two-increment, or three-increment development approach by specifying which capabilities to develop in each increment, whether to use the Services' POE or the CAIG ICE for AUPC estimates, and the EMD phase duration for each increment (Mortlock, 2020). All were from various defense acquisition career fields across the Army, Navy, and Air Force attending as students at Naval Postgraduate School in a Master of Science in Program Management. Responses came from active-duty officers and government civilians currently serving in the DoD acquisition workforce who hold various DAWIA acquisition certifications. None of the survey participants had prior experience working within either the JCM or JAGM programs of record. They completed the survey during coursework in a master's level academic program at Naval Postgraduate School. Before survey respondents completed the case study questions, they discussed the merits of critical thinking versus gut instinct reactions during decision making, risk vs. knowledge-based decision making and greater benefits of incremental development approaches.

The acquisition workforce represented by this sample includes approximately 150,000 total civilians (roughly 90 percent) and uniformed military members (roughly 10 percent) spanning 14 distinct career fields such as engineering; contracting; life cycle logistics; program management; production and quality management; test and evaluation; and business-financial management (Schwartz et al., 2016).

This study aggregates data from human respondents at the group level: frequency counts and percentages for each survey variable across the total number of respondents. For each decision factor the data records how many individuals chose each option across

choices made for the strategy option (single step vs. two-increment vs. three-increment choice), seeker type selection (single mode NDI TRL 9 vs. dual mode vs. tri-mode APB KPP TRL 6 choice), warhead type selection (single NDI TRL 9 vs. multipurpose APB KPP TRL 6 choice), propulsion selection (single motor NDI TRL 9 vs. common APB KPP TRL 6 choice), platform integration types (AH-64, AH-1, MH-60, F/A-18), schedule choice (48-month APB POE vs. 72/144-month ICE choice), and cost ($108K/$120K APB POE vs. $153K ICE choice). Frequency distributions for each decision variable provide human baseline data against which the study compares AI responses across all four layers of analysis.

2. AI Data Collection

The study gathered AI response data by presenting multiple artificial intelligence models with the same case study data and survey instrument under controlled conditions. Models were selected to form multi-tier comparison groups that isolate specific variables in the broader human-versus-AI comparison. Six commercial models were selected. These models have instruction-tuning and extensive reinforcement learning from human feedback (RLHF). The models selected include Claude Opus 4.6 and Claude Sonnet 4.6 (Anthropic), ChatGPT Instant 5.3 and ChatGPT Thinking 5.4 (OpenAI), and Gemini Pro and Gemini Thinking (Google). Two open-source models with instruction-tuning include Mistral Small 3 24B Instruct (Mistral AI, Paris) and Qwen 2.5 32B Instruct (Alibaba, China). Using this set of eight models creates analytical tiers that isolate specific variables in the human-versus-AI comparison.

Each model was prompted thirty times using an identical input to produce a distribution of responses. With a sample size of thirty observations per model, each AI sample produces its own response distribution that can be directly compared to the aggregated human distribution. Response variance within each model compared across all thirty runs measures internal consistency, while variance between models can be analyzed for inter-model agreement. Both aspects are required to fully answer this study's research questions on AI decision quality and reliability.

No information carried over from one run to the next. For commercial models this required initiating a new chat session for each run and utilizing the private / restricted

chat function, allowing for no data crossover. The entire prompt was input as one message, no follow-up prompting or asking the model to clarify was allowed. Anything the model produced in the first attempt at answering the survey was recorded. If the model asked a question instead of making a choice or refused to make a selection that behavior was captured in its entirety. Each run's output was recorded as delivered, with no modifications or retrospective cleaning. This process is critical for ensuring that each AI model is provided with only the information provided to human respondents, and that the output contains their unguided interpretation of that information.

The following input and output data were recorded for each run: a unique identifier for the run, date and time of the run, model version used, verbatim raw text response output, and coded responses for each survey variable extracted from output text and transformed to match the standardized data structure used for human responses.

3. AI Model Selection Rationale

This study selected AI models to build a structured comparison that isolates the impact of instruction tuning, RLHF alignment, training data sources, and model architecture on acquisition decision-making behavior. This methodology builds on the work of Csaszar, Ketkar, and Kim (2024), who compared human and AI-generated responses on a business strategy generation and evaluation task and concluded that LLMs have reached human-level performance on such tasks. However, as the authors discuss, one drawback of LLMs "is that [they] may replicate conventional strategies from their training data, rather than thinking 'outside-of-the-box' to develop novel plans" (Csaszar et al., 2024, Section 6). This limitation has direct implications on whether AI models can produce human-like trade-off analysis for acquisition planning. Models are separated into analytical tiers based on commercial versus open-source and whether the model has capability for reflexive reasoning.

Three commercial platforms were selected, each of which was tested in two variants to create a paired comparison: Claude (Anthropic) was selected as Opus 4.6 (capable of reasoning) and Sonnet 4.6 (standard version without reasoning capability), ChatGPT (OpenAI) was selected as Instant 5.3 (standard) and Thinking 5.4 (reasoning mode engaged), and Gemini (Google) was tested as Pro (standard) and Thinking

(reasoning mode). As open-source alternatives, the study also tested two instruction-tuned models: Mistral Small 3 24B Instruct (Mistral AI, Paris) and Qwen 2.5 32B Instruct (Alibaba, China). Mistral was chosen due to its use of an Apache 2.0 license and non-US based development team, signaling a different RLHF methodology than US-based commercial models. Qwen is included as it represents the most challenging test of data provenance impacts on acquisition decision-making: LLMs developed outside of Western contexts may produce dramatically different decisions based on their training data, which will contain a substantially different information distribution and cultural context than Western-produced models.

The eight models, presented in Table 1, represent the three most popular commercial families of large language models and leading open-source LLM alternatives available at the time of this research. Each model has been trained on large-scale internet datasets then refined through RLHF. While these models share this approach in common, different organizations developed them using different training datasets, RLHF methods, and engineering designs. As such, they serve a dual purpose of not only providing AI comparison group data to the human benchmark but also testing whether commercial implementation of RLHF produces significantly different outcomes when provided identical inputs.

Paired comparisons between the instant and thinking version within each commercial platform addresses a secondary research question: does how the AI models process information (instantly versus through deliberative reasoning) impact evidence of cognitive bias tendencies in acquisition decisions? If, for example, a model's thinking variant consistently produces different decisions than its standard variant there may be mitigating effects from the reasoning process itself.

Table 1. AI Model Selection Summary

Model	Tier	Analytical Role	Key Differentiator	Infrastructure
Claude (Anthropic) Opus 4.6 & Sonnet 4.6	Commercial	Primary AI comparison vs. human data	Constitutional AI alignment	Cloud API
Gemini (Google) Thinking & Fast	Commercial	Primary AI comparison vs. human data	Multimodal training pipeline	Cloud API
ChatGPT (OpenAI) 5.3 Thinking & Instant	Commercial	Primary AI comparison vs. human data	InstructGPT / RLHF lineage	Cloud API
Mistral Small 3 24B	Open-Source Instruct	Isolates instruction tuning vs. commercial RLHF	European-developed; different alignment approach	Local (RTX 3090, Q4_K_M)
Qwen 2.5 32B	Open-Source Instruct	Tests training data provenance effects	Chinese-developed; different data distribution	Local (RTX 3090, Q4_K_M)

4. Prompt Design and Standardization

To ensure valid comparisons between human and AI acquisition strategy selections, the researcher provided all models with the same information that human participants received. The prompt comprises three parts delivered to the model as a single message. The first part of the prompt ensures each model has the same context for taking on the assignment. The prompt tells the model that it is serving as a Program Manager for ACAT-1 programs and currently serves as the PM for the Joint Common Missile program. This introductory section matches the instructions provided to human respondents. The prompt then instructs the model to use only the program data provided to answer the survey questions, and that it should assume this is a live decision where the outcome is not known (to limit models from using external information about the current status of the JCM program).

The second part of the prompt contains the case study information. This includes the exact text from the Situation Overview, Background, Draft APB (including performance, schedule, and cost parameters), Work Breakdown Structure, Risk Assessment (including all 5 Risk items), and Technology Readiness Levels; as well as verbatim text from the CAIG ICE ($153mil). The researcher made no additions or subtractions to the case study data.

The third and final part of the prompt presents the eight survey questions. The prompt directs models to only use information provided in the prompt, choose from the options provided, and respond in a structure that matches the survey questions. Coding the output into discrete survey responses forces the AI models to make selections that can be analyzed similarly to how human survey data would be.

The study implemented several controls to reduce the risk of introducing bias into AI selections through the prompt design. First, the study used an identical prompt text across every model. This treats the model itself as the independent variable. Second, the prompt used no leading language, and neither hinted at nor suggested potential answers to the models. Third, the prompt maintained the order of the case study information as the survey presented it to human respondents. By controlling the order of information each model is given it ensures that any anchoring or ordering effects will impact human and AI respondents equally.

5. Rubric Development for Evaluation

In line with Chapter I, this study created an evaluation rubric to allow for consistent, standardized scoring across human and AI outputs. Recommendations from the GAO acquisition best practices literature (principally the knowledge-based acquisition practices identified in GAO annual weapon systems evaluations reports (GAO, 2018) and cost estimation best practices from the GAO Cost Estimating and Assessment Guide (GAO, 2020)) map to leading cognitive bias indicators from the behavioral decision-making literature (primarily the four bias constructs identified by Mortlock and Dew (2021) in defense acquisition and the anchoring/confirmation bias constructs from Tversky and Kahneman (1974)). The resulting rubric scores responses along five dimensions on a scale of 1 to 3 (1=Weak, 2=Adequate, 3=Strong), for a

maximum score of 15 across all five dimensions. Crucially, the rubric scores each of the five dimensions against only the data provided within the case study and survey instrument. The study neither expected nor permitted human or AI responders to access external information sources; the evaluation rubric scores how completely respondents engaged with and traded-off among the many competing data points contained within the case study document itself (including the Draft APB, the CAIG ICE, the risk assessment, the TRL information, and the CDD requirements). Table 2 depicts the full rubric.

Table 2. Evaluation Rubric for Acquisition Strategy Survey Responses

Dimension	Weak (1)	Adequate (2)	Strong (3)
1. Policy Conformance	KPPs omitted without explanation	Most KPPs addressed; relief incomplete	All KPPs addressed with relief plan
2. Analytical Rigor	Single data source only	Multiple sources; conflicts unreconciled	Multiple sources; conflicts reconciled
3. Internal Consistency	Multiple misalignments	One-two misalignments	Fully coherent package
4. Risk Integration	Risk data ignored	Overall risk acknowledged	Differentiated risk drives decisions
5. Cognitive Bias (inverse)	4+ indicators triggered	2-3 indicators triggered	0-1 indicators triggered
Composite	**5 (minimum)**	**10 (midpoint)**	**15 (maximum)**

Dimension 1: Policy Conformance (1-3): Selecting KPPs consistent with CDD requirements and proposing a major weapon system acquisition strategy meets thresholds for knowledge-based acquisition as defined by DoDI 5000.85 and is therefore prescriptive in GAO (2018) work on acquisition best practices. Score 3 (Strong): Response selects all threshold KPPs (tri-mode seeker, multi-purpose warhead, common motor, all four platforms) and either justifies development of all capabilities in one step or, if selecting an incremental approach, explicitly states how the risk associated with unverified KPPs will be relieved through a configuration steering board or CDD change process. There is recognition that selecting capabilities beyond the scope of the JROC-approved CDD requires affirmative relief. Score 2 (Adequate): Response fails to select all KPPs but explicitly addresses how deferred capabilities will receive JROC relief or selects all KPPs but does not comment on or recognize the programmatic risk of doing so given the Med/High integrated system risk rating. Score 1 (Weak): Response fails to

select threshold KPPs and does not provide any explanation or addressing KPP relief requirements or selects capability options that are not clearly defined by the CDD.

Dimension 2: Analytical Rigor (1-3): The GAO Cost Estimating and Assessment Guide recommends well-documented estimates that include comprehensive descriptions of data sources, underlying assumptions, and methodologies. Scoring in this dimension is based on evidence of data-driven reasoning across both selections and rationale (how closely did the respondent read and respond to the information provided within the case study?). Score 3 (Strong): Response references at least three internal case study data sources (the Draft APB schedule and cost, the CAIG ICE schedule and cost, the risk assessment ratings, and TRL data) and explicitly trade-off or reconcile the conflict between data sources. Example: Selecting the APB schedule but using the CAIG ICE's higher cost estimate with rationale that discusses why program-specific data points justify departure from the independent estimate. Each selection for both schedule and cost rationale includes at least one program-office-aligned justification and one ICE-aligned justification. Score 2 (Adequate): Response mentions at least two internal data sources provided within the case study but fails to explicitly acknowledge conflict between data sources. Example: The response acknowledges the CAIG ICE exists as part of the case study data but selects APB values without discussing why the independent estimate was not used. Score 1 (Weak): Response only references a single internal data source (usually the APB or CDD) and makes selections without regard to or mention of other contradictory information provided by the CAIG ICE risk assessment contained within the case study data. All rationale selections are APB-aligned and do not include any of the CAIG ICE or risk assessment rationale options provided in the survey instrument.

Dimension 3: Internal Consistency (1-3): Strategy selections that are logically coherent or self-reinforcing meet the GAO's (2018) definition of an informed commitment and are also operationalized as sound acquisition planning in the literature (Drezner & Krop, 1997). Scoring in this dimension assesses how well performance, schedule, cost, rationale, and importance ratings support each other. Score 3 (Strong): Performance selections, schedule/cost, and rationale form a coherent package. If all KPPs are selected (highest size/complexity option), either schedule and cost match the CAIG ICE ranges or the response includes program-specific data used to justify selection of

APB estimates despite the independent estimate. Importance ratings are not contradicted by actual selections. Factors the respondent identified as significantly important are present in the strategy selections they made. Score 2 (Adequate): All selections are consistent except for one glaring inconsistency between importance rating and selection or rationale selection and actual strategy choice. Example: Selects 48-month schedule but rated the risk assessment as significantly important. Or cites the CAIG ICE as a rationale selection but chooses the APB cost. Score 1 (Weak): There are multiple inconsistencies between what the respondent said was important and their actual selections. Example: selecting all KPPs with a 48-month schedule and $108K AUPC while ignoring the CAIG ICE's 72–144-month schedule estimate and $153K cost estimate. Or rating factors as significantly important that contradict the actual strategy selections made (e.g., rates supporting furtherance of multi-mission F-16 capabilities as not important but selects no aircraft that require the multi-purpose warhead capability). Selecting all KPPs with the shortest schedule and lowest cost represents a mutually unrestricted set of commitments given the program characteristics established by the case study information.

Dimension 4: Risk Integration (1-3): The GAO (2018) found programs that completed knowledge-based acquisition activities, including conducting risk-informed technology demonstrations, realized 56 to 63 percentage points lower cost growth than programs that did not conduct KBAs. Scoring responses based on whether their selections adequately incorporate the risk information into their strategy choice. Score 3 (Strong): Response differentiates medium vs. medium/high risk WBS elements (seeker and motor vs. warhead and integration) and uses the risk ratings to inform either capability sequencing, schedule, or overall strategy selection. If incremental approach is selected, higher-risk capabilities are deferred to later increments or given additional schedule margin. Score 2 (Adequate): Respondent recognizes the Med/High integrated system risk rating from the Case Study and changes either schedule or cost estimate (adds schedule margin or cost) for at least one program component but does not differentiate between component-level risk profiles. Score 1 (Weak): Respondent ignores risk rating and selects most aggressive schedule and lowest possible cost without adjustment OR rates risk assessment as significantly important to contractors but makes selections that

ignore its implications. There is no evidence that the granular risk data for each WBS element impacted strategy choices.

Dimension 5: Cognitive Bias Indicators (1-3) Scoring responses against the 6 cognitive bias attributes and constructs identified in the literature review. Responses that trigger more bias criteria will receive lower scores on this dimension. Scoring in this dimension is inverse (fewer bias criteria triggered equals higher score). Score 3 (Strong / Low Bias): 0 or 1 bias attribute triggered. Response must adjust away from the APB anchor toward the CAIG ICE for either schedule or cost values, make difficult capability tradeoffs informed by risk data, use both supporting and contradictory information in rationale selections, and provide importance ratings that match actual selections. Score 2 (Moderate Bias): 2 or 3 bias attributes triggered. Response evidence of anchoring to APB values for both schedule and cost and/or evidence of the planning fallacy (making inside view justification without outside view adjustment) but also makes some effort to adjust from APB values or make trade-offs. Score 1 (High Bias): 4–6 bias attributes triggered. Response must select APB values for both schedule and cost with all KPP capabilities selected despite both CAIG ICE values and risk assessment data (optimism bias and anchoring), make no difficult trade-offs (difficulty making tradeoffs), select only rationales that support the selected strategy and ignore available rationales that would contradict the choice (confirmation bias), and provide importance ratings that contradict the actual selections made (planning fallacy).

The following criteria have been developed to operationally define each cognitive bias attribute within the scoring rubric.

Optimism Bias: Response selected 48-month schedule AND $108k or $120k AUPC cost AND did not select CAIG ICE rationale AND did not cite risk assessment as a driver of schedule or cost.

Anchoring: Response selected 48-month schedule AND $108k AUPC cost with no adjustment to CAIG ICE values.

Planning Fallacy: Response selected the 48-month schedule AND rated either TRL maturity or the risk assessment as significantly important without selecting any of the CAIG ICE rationales (inside view bias).

Difficulty Making Trade-offs: Response selected all KPPs across all component AND platforms with no cost or schedule adjustments for increased capability scope despite the Med/high risk assessment (replicates behavioral pattern observed by Mortlock and Dew (2021) in the original JCM program).

Confirmation Bias: Response selected only rationales that support the selected strategy without selecting any rationales that would contradict the chosen strategy (e.g., selects APB cost and schedule values but does not select CAIG ICE rationale as a justification).

Legacy Preference: Response selects COTS components for seeker and motor without discussing either familiarity or proven past performance as part of their justification.

6. Rubric Application and Scoring Summary

Each human respondent and each AI model run receives a score on all five dimensions. Composite scores range from 5 (weakest) to 15 (strongest). For AI models with thirty runs each, the analysis calculates the mean composite score and standard deviation to enable comparison with the human respondent distribution. The rubric is applied identically to human and AI data to ensure comparability.

The same evaluator rated the entire 240 sets of AI output and used the rubric to evaluate the human aggregate frequency distributions. The three limitations of using a single evaluator are mitigated by the fact that all evaluation decisions can be reduced to verifiable checks of whether specific survey boxes were checked instead of subjective qualitative judgments. For instance, the Optimism Bias indicator triggers based solely upon the selection of all four of the following: the 48-month timeline, one of either the APB Cost Estimates of $108K or $120K, No CAIG ICE Rationale, and No Risk-Driven Rationale. There is therefore no way for an evaluator to interpret how to score these items. Secondly, the evaluator re-evaluated a random sample of 24 runs from the 8 AI Models (at least one incremental run from each model's set of runs) after a three-week period to establish intra-rater reliability. The percent agreement was 89.2%, with Cohen's Kappa = 0.80. This indicates substantial agreement (Landis & Koch, 1977). Because the

human respondent data Mortlock (2020) collected consisted of aggregated frequency distributions rather than individual-level survey responses, calculating intra-rater reliability for the human data was not possible. Given that the rubric decision rules are identical across both datasets and result solely in checkbox evaluations, the reliability of the AI sample data represents strong evidence of consistent rule application.

C. COMPARISON FRAMEWORK AND STATISTICAL METHODS

The comparison of human and artificial intelligence (AI) generated data follows a four-layered analytical approach in order to answer the three research questions at increasingly deeper levels of analysis. The first layer presents an objective, factual account of what each selection group chose across all survey items. For each decision item, the analysis reports the frequency distributions for the human sample population and for each run of each of the AI models. Thus, this layer answers the secondary research question at its most basic point: How do acquisition decisions generated by AI differ from those made by professional humans?

The second layer addresses the issue of whether each selection represents a coherent, internally consistent set of decisions. Four distinct consistency metrics are employed in assessing selections. The KPP Alignment Score evaluates how well performance selections are aligned with the Key Performance Parameters of the CDD. Schedule Cost Performance Coherence assesses the realism of selected schedules and costs relative to selected performance configurations. The Rationale Decision Alignment metric examines the logic-based rationales supporting selections. Finally, the Risk Decision Integration metric evaluates how well selections take into consideration the results of the risk assessments.

The third layer transforms the cognitive bias constructs identified in the literature review into specific measurable indicators. Six biases were examined. When respondents choose a schedule and/or cost value less than or equal to their internal APB estimate(s) but have countervailing evidence in the form of the CAIG ICE and/or the Medium-High risk assessment, optimism bias is detected. The analysis detects anchoring by examining which data source exerts the strongest pull on responses. The planning fallacy occurs when respondents select the most aggressive possible schedule, rate the risk assessment

as important, yet discount the importance of the CAIG ICE. Difficulty making tradeoffs occurs when respondents relax all three elements of the triple constraint rather than holding two constant and adjusting one. Confirmation bias occurs when rationale selections show systematic favoritism towards confirming strategies selected versus ignoring/omitting contradicting evidence. Legacy preference occurs when respondents select COTS components primarily due to familiarity instead of being driven by data. Each bias indicator is applied equally to both human and AI data.

The fourth layer compares the different AI models to each other. Inter-model agreement refers to the extent to which different models agree on the same decisions. Intra-model consistency refers to the variability among each model's thirty responses. Divergence between trained and untrained models compare instruction-tuned model responses to any base model responses to see if there was a systematic shift in acquisition decisions based upon alignment training.

Since Strategy choice, performance selections, and rationale selections are categorical variables, comparisons between human and AI data will be completed via chi-square tests of independence, or Fischer's exact test when expected cell counts are small. Schedule and cost selections can be thought of as binary classifications (APB vs. ICE), these will also be analyzed via Fisher's Exact Tests. The Likert scale has only three possible values, and the importance ratings on a three-point Likert scale will be evaluated using Two-Proportion Z-Tests. Reference values will be determined using the objective program data: APB values and CAIG ICE values will represent reference points for anchoring; CDD KPP Requirements will serve as performance baselines; and risk assessment ratings will serve as standards for informed decisions based on risk.

This study includes effect size along with every significance test to report on the practical impact of the results, regardless of sample size. The p-value alone may be misleading: with a total of 271 observations (31 human and 240 AI), even minor differences will likely result in a statistically significant result. When reporting effect sizes, they eliminate the influence of sample size so that you see if the difference is practically large enough to matter. In addition, this study uses the following measures of effect size: Cramer's V for multi-category Chi-Square tests, Cohen's h for pairwise

proportion comparison. To provide conventional benchmarks, the following was interpreted as: Small Effects V or h ≈ 0.10 – 0.20, Medium Effects ≈ 0.30 – 0.50, Large Effects > 0.50 for Cramer's V, > 0.80 for Cohen's h. By providing these values it ensures that only those differences that are both statistically significant and have practical implications support the conclusion of this study.

In addition, this study calculates the Shannon Entropy (H) for each group's strategy selection distribution to quantitatively evaluate response diversity. Entropy evaluates the dispersion of a distribution using a single number. An entropy value of zero indicates that all respondents selected the same option, demonstrating complete uniformity. A maximum entropy value suggests that selections were distributed equally across all available options. With three strategy categories, a maximum entropy value would equal H_max = log2(3) = 1.585 bits. As such, this measure allows direct examination of whether AI has eliminated useful decision-making variation through collapsing into a single strategy, an issue central to this research.

Since the human respondent data collected by Mortlock (2020) was provided in aggregate format (i.e., frequencies of strategies, schedules, and costs selected), and not as individual level survey responses, the inferential testing used in this research is based upon grouped frequency totals (the total number of respondents who selected each strategy, schedule, etc.). The statistical methods to test hypotheses used in this study; Chi-Square Tests, Fisher's Exact Test(s), Two-Proportion Z-test(s), and Shannon Entropy, do not rely on individual level data; they only require grouped frequency totals. Grouped frequency totals support the inferential tests listed above (chi-square, Fisher's exact, two-proportion z-tests, and Shannon entropy), which compare categorical selection patterns between groups. However, grouped frequency totals do not support inferential comparison of rubric composite scores, which require individual-level data from both groups. This is a methodological limitation. Individual level human response data could have allowed for additional non-parametric statistical tests to compare rubric composite scores of groups (Mann Whitney U test). Therefore, future studies collecting individual level human response data will allow for extension of the inferential testing beyond what grouped frequency totals can provide. The study set significance thresholds for all tests at $\alpha = 0.05$.

D. RESULTS

This section presents the findings from both human and AI respondents across all four layers of analysis. The human baseline results are presented first, followed by AI results, to establish the benchmark against which all AI-generated decisions are measured.

1. Human Respondent Results

Mortlock (2020) obtained the human respondent answers used for comparison (presented in Figure 3) through primary data collection. 22.6 percent of respondents elected for the single-step strategy; 41.9 percent elected for the two-increment strategy and 35.5 percent elected for the three-increment strategy. These results confirmed Mortlock's (2020) primary finding that acquisition professionals use knowledge of TRLs and risk ratings to recommend strategy components but have difficulty prioritizing among the elements of the triple constraint. As Mortlock (2020) noted, the results provided evidence that professionals tended to relax cost, schedule, and performance simultaneously rather than selecting one to adjust, putting PMs in the difficult position of not being able to deliver on any constraint while increasing the risk that the program would not receive approval at the milestone. Over three quarters (77.4 percent) opted to develop JCM incrementally rather than select the single-step strategy that was eventually approved for JCM. Zero percent proposed a strategy similar to the one ultimately pursued by the Services over ten years later (two-way seeker, NDI warhead, NDI motor, integration on AH-64 and AH-1 only in first increment).

		Seeker			Warhead		Propulsion		Platform				Schedule (EMD length)		Cost (AUPC)	
	Respondents (n)	Single Mode (NDI) TRL 9	Dual Mode	Tri-mode APB KPP TRL 6 Med Risk	Single (NDI) TRL 9	Multipurpose APB KPP TRL 6 Med/High Risk	Single motor (NDI) TRL 9	Common APB KPP TRL 6 Med Risk	AH64 APB KPP	AH1 APB KPP	MH60 APB KPP	F18 APB KPP	48 months APB POE	72 or 144 months ICE	$108K or $120K APB POE	$153K ICE
	31															
Single Step	7		1	6	1	6	1	6	6	6	6	7	1	6	2	5
		Two-Increment Approach								**Two-Increment Approach**						
Increment I	13		8	5	7	6	3	10	12	11	10	5	7	5	8	4
Increment II				13		13		13	13	13	13	13	3	8	5	8
		Three-Increment Approach								**Three-Increment Approach**						
Increment I		4	5	2	8	3	10	1	10	8	6	5	9	2	7	4
Increment II	11		4	7	5	6	8	3	10	9	9	8	7	4	6	5
Increment III				11		11	1	10	10	9	9	10	7	4	6	5

Figure 3. Survey Data Results – Human Respondents (n=31). Source: Mortlock (2020, p. 292)

Among respondents choosing the single-step strategy there was strong agreement on selecting full KPP performance while the plurality selected the longer ICE schedule and higher ICE AUPC. Among respondents selecting the two-increment strategy, the seeker was the most common capability to defer while the F/A-18 Super Hornet was the platform most commonly deferred to a later increment. Among respondents selecting the three-increment strategy, first increments were the most common across all strategies with most respondents deferring the tri-mode seeker, multipurpose warhead, and common motor to subsequent increments while holding both the APB schedule and cost estimate constant for the first increment.

Several notable trends crossed strategy selections and have direct bearing on the cognitive bias analysis. Across all respondents, only 45 percent elected to hold constant the approved Service cost and schedule estimates while making their programs capable over time; the remaining respondents relaxed all three legs of the triple constraint rather than holding two fixed and trading off between the third and one decision variable, indicating a possible susceptibility to difficulty-in-making-tradeoffs bias. Respondents did not consistently use their capability deferral decisions with the technology readiness and risk information provided; while the seeker (medium risk, TRL 6) and the warhead (medium/high risk, TRL 6) were close, the warhead was rated significantly higher on the risk spectrum but was deferred less frequently than the seeker. Finally, across both the two-increment and three-increment strategies, respondents most often chose to defer the F/A-18 platform, correctly recognizing that the eventual JAGM program would ultimately drop the F/A-18 as a threshold platform, but arrived at that conclusion through a mix of rationale rather than uniformly applying the data to reach their decision. The significant variance in answers provided by respondents across nearly every decision node highlights the key finding from the exercise: when provided with the same programmatic information, acquisition professionals will recommend vastly different acquisition strategies.

2. AI Respondent Results

AI data collection ran the same standardized prompt across eight AIs thirty times each for a total of 240 AI runs and used no persona variation, no emphasis framing, and

no temperature changes. The eight models included Claude Opus 4.6, Claude Sonnet 4.6, ChatGPT Instant 5.3, ChatGPT Thinking 5.4, Gemini Pro, Gemini Thinking, Mistral Small 3 24B Instruct, and Qwen 2.5 32B Instruct.

a. Strategy Selection: Near Universal Convergence

The most significant result from the AI data collection is the overwhelming preference for Strategy A, the Single Step Acquisition Strategy. Seven of eight models chose Strategy A thirty times out of thirty runs (100 percent): Claude Opus 4.6, ChatGPT Instant 5.3, ChatGPT Thinking 5.4, Gemini Pro, Gemini Thinking, Mistral Small 3 24B Instruct, and Qwen 2.5 32B Instruct. The study observed consistent results across the three commercial providers (both reasoning and standard models), and both open-source models. Between them, these seven models account for 210 uniform selections of Strategy A out of 240 possible AI runs.

Claude Sonnet 4.6 was the only model to produce a different distribution of strategy selections. Out of thirty runs, Claude Sonnet chose Strategy A 21 times (70 percent) and Strategy B (Two-Increment) 9 times (30 percent). Claude Sonnet did not produce any runs selecting Strategy C. Claude Sonnet's divergence from the other seven models is noteworthy in its own right: two instruction-tuned models developed by the same company (Anthropic) produced substantially different results based on the model's size and capability. Sonnet's 70 percent Strategy A selection rate was notably lower than the 100 percent rate produced by Opus. As the analysis shows, this indicates that acquisition decision-making may be sensitive to a wider range of factors than just dataset or alignment technique and it may also depend on model architecture. Additionally, notable is that Claude Sonnet was the only model to recommend any form of incrementality. Across its nine selections of Strategy B, Claude Sonnet selected a phased capability approach to increments. Increment I consistently featured a dual-mode seeker (8/9), single COTS warhead (6/9), and common motor (9/9), and was paired with delaying full-tier-count platforms AH-64 and AH-1Z until Increment II. Multiple models across all generations included these same programs on Increment I alongside the full tri-mode seeker, multi-purpose warhead, and other platforms A notable point of comparison is how human survey respondents answered the same question. While just 3.8 percent of

AI runs selected either Strategy B or C, 77.4 percent of humans selected those same strategies. Where human respondents overwhelmingly opted for an incremental approach, AI respondents did the opposite and favored the Single Step approach. The human and AI results here are nearly the inverse of each other. Figure 4 visualizes this distribution.

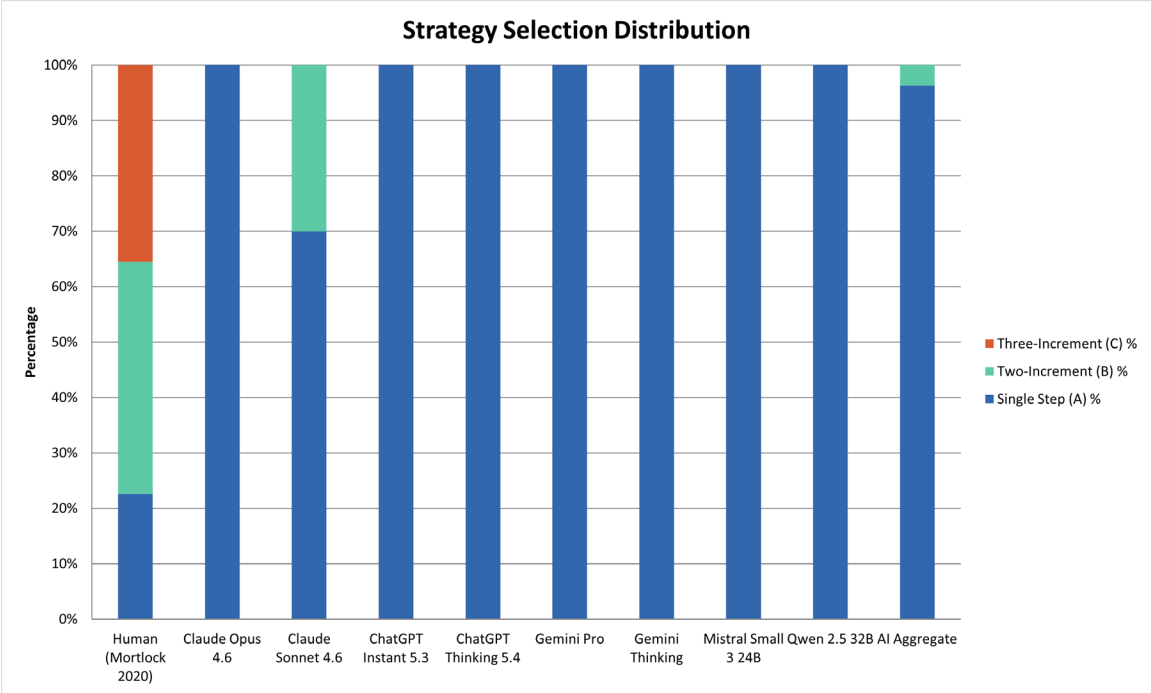

Figure 4. Strategy Selection Distribution

Table 3. Strategy Selection Distribution

Source	n	Single Step (A) %	Two-Increment (B) %	Three-Increment (C) %
Human (Mortlock 2020)	**31**	**22.6%**	**41.9%**	**35.5%**
Claude Opus 4.6	30	100.0%	0.0%	0.0%
Claude Sonnet 4.6	30	70.0%	30.0%	0.0%
ChatGPT Instant 5.3	30	100.0%	0.0%	0.0%
ChatGPT Thinking 5.4	30	100.0%	0.0%	0.0%
Gemini Pro	30	100.0%	0.0%	0.0%
Gemini Thinking	30	100.0%	0.0%	0.0%
Mistral Small 3 24B	30	100.0%	0.0%	0.0%
Qwen 2.5 32B	30	100.0%	0.0%	0.0%
AI Aggregate	**240**	**96.3%**	**3.8%**	**0.0%**

A descriptive comparison of strategy selections indicates there is a dramatic difference in the way humans and AI select strategies, however, descriptive statistics

cannot establish if the differences are due to chance or a true relationship. This question can be answered through a chi-squared test of independence. The chi-squared test examines if the strategy selection patterns depend upon the type of respondent (AI vs. Human) or if they operate independently. The analysis conducted a chi-squared test using the 3×2 contingency table (strategy A/B/C × Human/Aggregate AI). The chi-squared test revealed a statistically significant association between respondent type (Human vs. AI) and the pattern of strategy selections χ^2 (2) = 151.44, $p < .001$. Since one of the expected cell counts (1.26) does not meet the chi-square assumption requirement of being greater than 5, the analysis used a Fisher's exact test to validate the findings. Fisher's exact test provides an exact probability without utilizing the large sample approximation. The results also validated that the association exists ($p < .001$). Cramer's V is a standardized measure of effect size and ranges from 0 to 1. The Cramer's V measured in this study resulted in a value of 0.748. A large effect would occur when Cramer's V is equal to or larger than 0.50. Therefore, we can conclude that respondent type (Human vs. AI) accounts for a substantial portion of the variance in strategy selections and the observed difference is not due to statistical chance.

The next step is to collapse strategy selections into a two-category comparison (Single-step vs. Incremental) and further highlight the differences between human and AI strategy selections. Fisher's exact test confirmed that human and AI strategy selections were divergent as well as highly disparate, OR = 0.011, $p < .001$, Cohen's h = 1.76. Cohen's h is a specific measure of effect size for proportion comparisons, and a large effect is typically defined as $h > 0.80$. In this case Cohen's h equaled 1.76, exceeding the large-effect threshold by more than a factor of two, resulting in an extremely large practical difference between human and AI strategy preferences. A pairwise Fisher's exact test was performed to assess each individual AI model against the human baseline. All eight models demonstrated a significant difference from humans (all $p < .001$; Cohen's h ranged from 0.99 for Claude Sonnet to 2.15 for the seven uniform models). Table 4 displays these pairwise results.

Table 4. Pairwise Fisher's Exact Tests – Human vs. Each AI Model

Model	n	Single-Step %	OR	p-value	Cohen's h	Effect
Human (Mortlock, 2020)	31	22.6	-	-	-	Base
Claude Opus 4.6	30	100.0%	0.000 (∞)	< .001	2.151	Large
Claude Sonnet 4.6	30	70.0%	0.125	< .001	0.992	Large
ChatGPT Instant 5.3	30	100.0%	0.000 (∞)	< .001	2.151	Large
ChatGPT Thinking 5.4	30	100.0%	0.000 (∞)	< .001	2.151	Large
Gemini Thinking	30	100.0%	0.000 (∞)	< .001	2.151	Large
Gemini Pro	30	100.0%	0.000 (∞)	< .001	2.151	Large
Mistral Small 3 24B	30	100.0%	0.000 (∞)	< .001	2.151	Large
Qwen 2.5 32B	30	100.0%	0.000 (∞)	< .001	2.151	Large

Note. Each test compares the model's Single-Step selection rate against the human baseline rate of 22.6% using a 2×2 Fisher's exact test (Single-Step vs. Incremental × Human vs. AI Model). OR = odds ratio; ∞ indicates zero incremental selections in the model (perfect separation, OR undefined). Cohen's h thresholds: small = 0.20, medium = 0.50, large = 0.80. All eight models differ significantly from humans at $p < .001$.

To quantify the imbalance in diversity between human and AI strategy selections Shannon entropy provides a singular metric to represent the degree of dispersion of a groups' selections across available options. When measuring entropy, values range from 0 to 1.585, where 0 represents all members selecting the same option and 1.585 represents perfect evenness, with each member splitting their selections perfectly across all available options. Human respondents achieved nearly perfect evenness, with an average entropy value of H = 1.54 bits. This represents approximately 97.2% of maximum possible diversity. The seven uniform models registered 0 entropy, all selecting the same strategy, resulting in absolute uniformity. Only Claude Sonnet produced moderate diversity with an entropy value of H = 0.88 bits, representing approximately 55.6% of maximum possible diversity. The overall average entropy value for the combined AI responses was very low with an average value of H = 0.23 bits, approximately 14.6% of maximum possible diversity. This supports the idea that AI models did not simply choose alternative strategies to those chosen by humans but instead eliminated all forms of strategic variability.

b. Performance Selections: Full KPP Retention

For every model and for nearly every run across all models, AI responses selected a full KPP performance selection in every category: Tri-Mode Seeker (KPP), Multi-Purpose Warhead (KPP), Common Motor (KPP), AH-64 platform thresholding, AH-1Z platform thresholding, F/A-18 platform thresholding, MH-60 platform thresholding. AI systems overwhelmingly selected the most technically aggressive, highest-risk set of performance parameters available.

Incremental human respondents deferred capability selections at much higher rates. Across all categories, about 71 percent of incremental respondents deferred the seeker, 63 percent deferred the warhead, 54 percent deferred the propulsion motor, and about half deferred the inclusion of the F/A-18 platform. Human respondents exhibited measurable trade-off behavior. AI models did not. Table 4 and Figure 5 show these rates of capability deferral side-by-side. The left set of bars charts the percentage of human incremental respondents that deferred each major capability element to a future increment. The right data point measures deferral across all AI models and is nearing zero percent. The divergence between human incremental respondents and AI models on this key decision point is among the most substantial results of this research.

Table 5. Capability Deferral Rates

Capability Element	Human Incremental (% deferred)	AI Aggregate (% deferred)
Tri-mode Seeker	70.8%	3.7%
Multipurpose Warhead	62.5%	3.7%
Common Motor	54.2%	3.7%
F/A-18 Platform	50.0%	3.7%

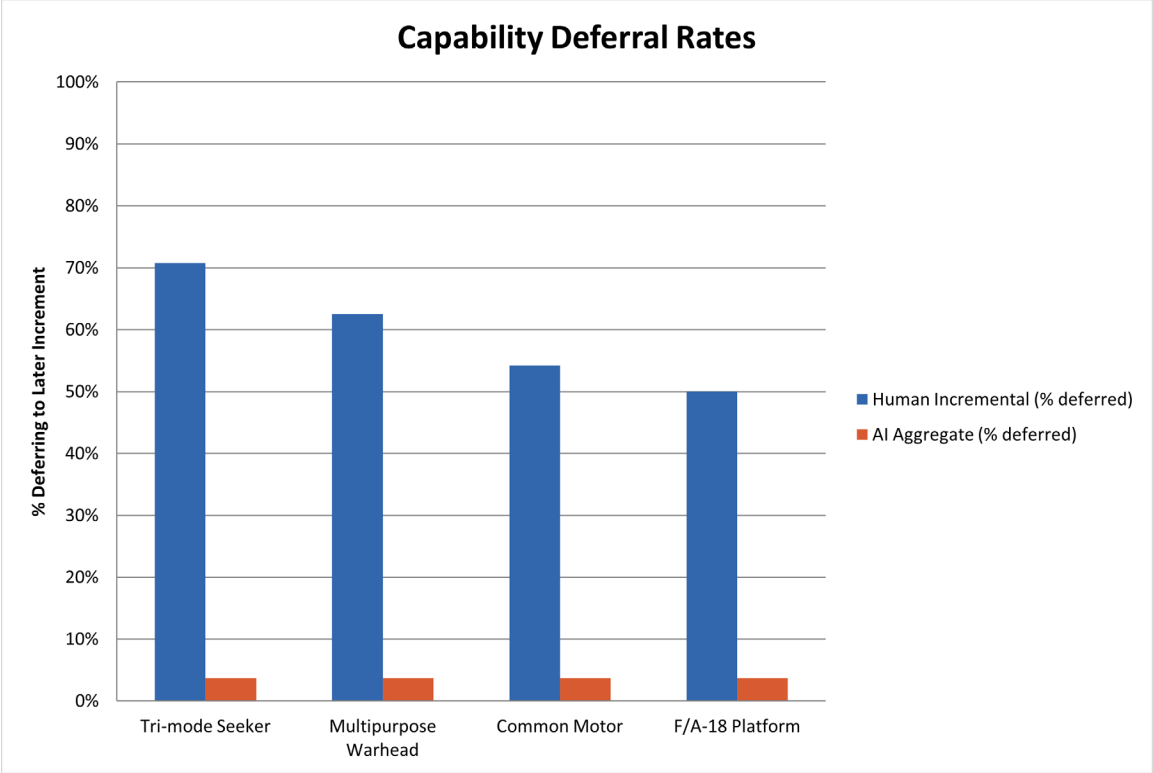

Figure 5. Capability Deferral Rates

In comparison to seeking deferrals, there are no statistically significant differences between the incremental respondent data sets (i.e., the 24 human and 9 Claude Sonnet incremental data points) for the overall seeker deferral rate ($p = .394$; 70.8% [human] vs. 88.9% [Sonnet]) or for the warhead deferral rates ($p = 1.000$; 62.5% [human] vs. 66.7% [Sonnet]). There were, however, two very distinct deviations from this trend. The first is that common motor deferral was a dramatically different experience with respect to the human incremental respondent population (54.2%) as opposed to the incremental Sonnet runs (0%; $p = .005$, Cohen's $h = 1.65$, large effect). This is in direct contrast to the second deviation where Sonnet deferred the F/A-18 platform in all (100%) of its incremental runs and the human incremental respondent population deferred only 58.3% of the time ($p = .032$, Cohen's $h = 1.40$, large effect).

c. Schedule and Cost Selections: Anchoring to the APB

All AI models chose the most aggressive schedule available, 48 months for EMD, aligning with the Draft APB's based on the CDD IOC and POM funding position. Human respondents also selected the 48-month schedule in approximately half of all

cases. 7 out of 8 AI models selected the 48-month schedule and APB-aligned cost in 100 percent of their 30 runs each. As noted previously, Claude Sonnet 4.6 varied from the other models consistently enough to be considered a distinct pattern rather than random noise.

Claude Sonnet ran 21 single-step selections and only 6 chose the 48-month APB schedule; the other 15 selected the longer CAIG ICE schedule. Similarly, only 6 of Claude Sonnet's single-step runs selected the APB cost while 15 chose the $153K CAIG ICE cost. This means Claude Sonnet was the only model in this study that exhibited any statistically significant sensitivity to CAIG ICE inputs, even among its single step runs. Across Claude Sonnet's 9 two-increment runs, its Increment I schedule choices broke evenly at 3 APB and 6 ICE. However, Increment II schedule choices went the opposite direction, breaking at 5 APB to 4 ICE. Interestingly, Claude Sonnet's Increment II cost choices included some runs that selected the $153K CAIG cost estimate.

The AI models have again behaved in uniformly aggregative ways that allow analysis across all models. Table 6, Figure 5, and 6 presents each model's schedule and cost selections relative to the source they align with. The APB estimate was clearly optimistic relative to the independent CAIG ICE estimate; however, every AI model except Claude Sonnet selected the APB schedule and cost every single time. Human respondents split their selections between the two options on each question. Claude Sonnet opted for the CAIG ICE in slight majority across both questions, showing a clear sensitivity to CAIG inputs that the remaining seven models do not. This uniform pattern shows evidence for anchoring: the draft APB provided schedule and cost values that were the first, most authoritative numbers in the case study data. The data provided the CAIG ICE values later as a contrast value. Although an independent group generated the CAIG ICE based on analogous historical programs, the models anchored to the APB values and adjusted zero.

Table 6. Schedule and Cost Anchoring – APB vs. CAIG ICE Selection Rates

Source	Schedule: 48mo APB %	Schedule: CAIG ICE %	Cost: APB ($108-120K) %	Cost: CAIG ICE ($153K) %
Human	42.0%	58.0%	55.0%	45.0%
Claude Opus 4.6	100.0%	0.0%	100.0%	0.0%
Claude Sonnet 4.6	20.0%	80.0%	37.0%	63.0%
ChatGPT Instant 5.3	100.0%	0.0%	100.0%	0.0%
ChatGPT Thinking 5.4	100.0%	0.0%	100.0%	0.0%
Gemini Pro	100.0%	0.0%	100.0%	0.0%
Gemini Thinking	100.0%	0.0%	100.0%	0.0%
Mistral Small 3 24B	100.0%	0.0%	100.0%	0.0%
Qwen 2.5 32B	100.0%	0.0%	100.0%	0.0%

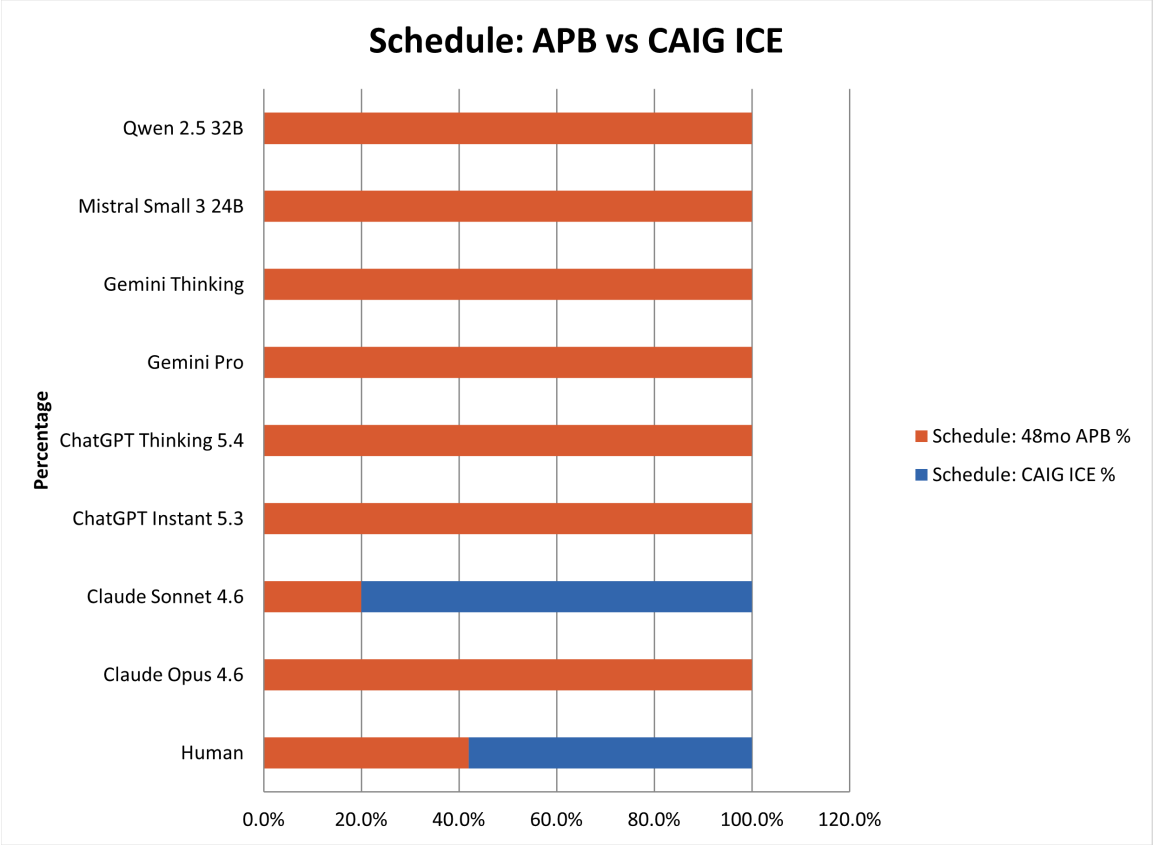

Figure 6. Schedule Anchoring

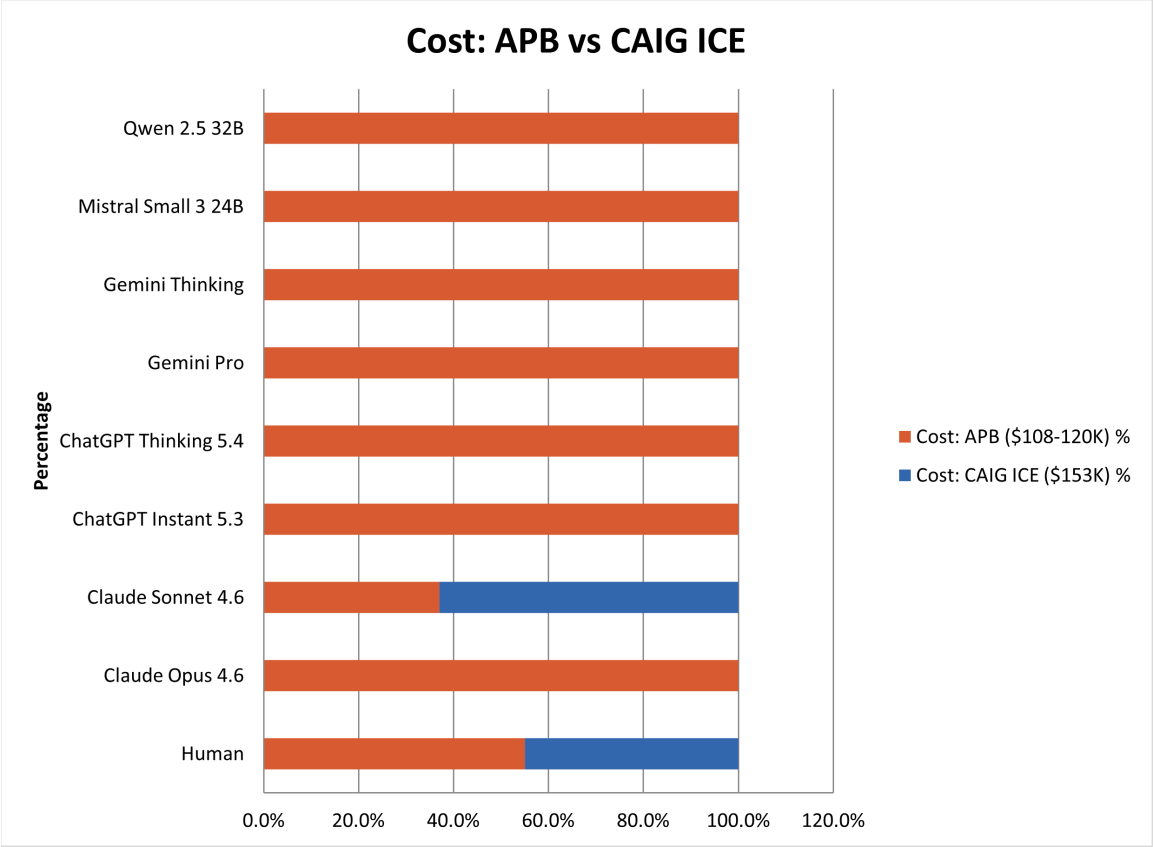

Figure 7. Cost Anchoring

Fisher's Exact Tests determine whether the observed differences in Schedule and Cost Anchoring choices reached statistical significance. On the schedule dimension, 56.7% of Humans selected the APB 48-month estimate as their Primary Increment Level, whereas all eight AI models selected this option 91.2% of the time (OR = 0.125; $p < .001$; Cohen's H = .84; Large Effect). On the cost dimension, 56.7% of Humans selected an APB estimate for total costs, while the same AI models selected this option 92.1% of the time (OR = 0.112; $p < .001$; Cohen's H = .87; Large Effect). These results indicate that both Humans and AI models anchored their responses to Program Office Estimates; however, the AI models did so at a statistically significant and substantially higher rate than Humans.

Claude Sonnet did not make cost selections that differed significantly from those made by Humans (APB = 36.7% versus APB = 56.7%), Fisher's Exact p = .195; Cohen's H = .40 (Small-Medium Effect). Similarly, although Claude Sonnet's schedule selections (APB = 30.0%) approached significance relative to Human selections (p = .067; Cohen's

H = .55; Medium Effect), they did not achieve significance. These results place Claude Sonnet in a behavioral space intermediate between Humans and the other seven AI models examined. Although Claude Sonnet demonstrated some of the same strategic formulations as the other AI models, it exhibited a degree of responsiveness to independent cost and schedule estimates similar to that exhibited by Humans.

d. Rationale and Importance Ratings

AI model selections for the rationale questions were extremely consistent. There were seven selectable rationale options for schedule and cost decisions. Those seven options came verbatim from the competing schedule and cost data inputs provided to respondents within the case study data. The most common schedule rationales selected were: "supports the warfighter required IOC," "specified in draft APB," "supported in Service POM funding positions," "supported by the risk assessment or TRL levels," and "supported by the performance capability development strategy." The two rationales that referenced the CAIG ICE directly ("Supported by the low CAIG ICE analysis" and "Supported by the high CAIG ICE analysis") appeared as choices in the same dropdown menu. While they were present in the same menu, AI models almost never selected them. As was the case with selected schedule values, AI models built justifications based around the case study data that supported the decisions they made and ignored the major dissenting voice in the data set, even though it was offered as a selectable option in the question stem.

When forced to rate the importance of various input factors, AI models almost always chose "Significantly Important." The lone exception was typically the JCM Risk Assessment, which AI models consistently rated as only "Moderately Important" despite posing the most significant downside risk to selecting a 48-month single-step acquisition strategy.

E. COGNITIVE BIAS ANALYSIS: AI VS. HUMAN COMPARISON

The six-construct cognitive bias analytic rubric is mapped in Figure 8. Counter to expectation that AI respondents would show less cognitive bias than humans, each of the AI models triggered indicators of four or five constructs of bias. The process that

generated these indicators is not the psychological process that drives cognitive bias in humans. This finding is not necessarily at odds with Alon-Barkat and Busuioc's (2023) warning that AI might impose new biases into human–algorithm relations, nor with more recent work on trustworthy AI for decision-making, which notes that "data-driven AI can inherit biases from historical data, leading to biased or unfair outcomes." (Miedema et al., 2026, p. 7). However, this study shows a type of AI bias not yet discussed in acquisition literature. This is a structural incentive for AI models to overweigh data points simply because they number more than alternatives. Such systematic data-weighting bias causes AI models to optimize toward the answer that most data points support and to underweight contrarian evidence no matter how analytically sound.

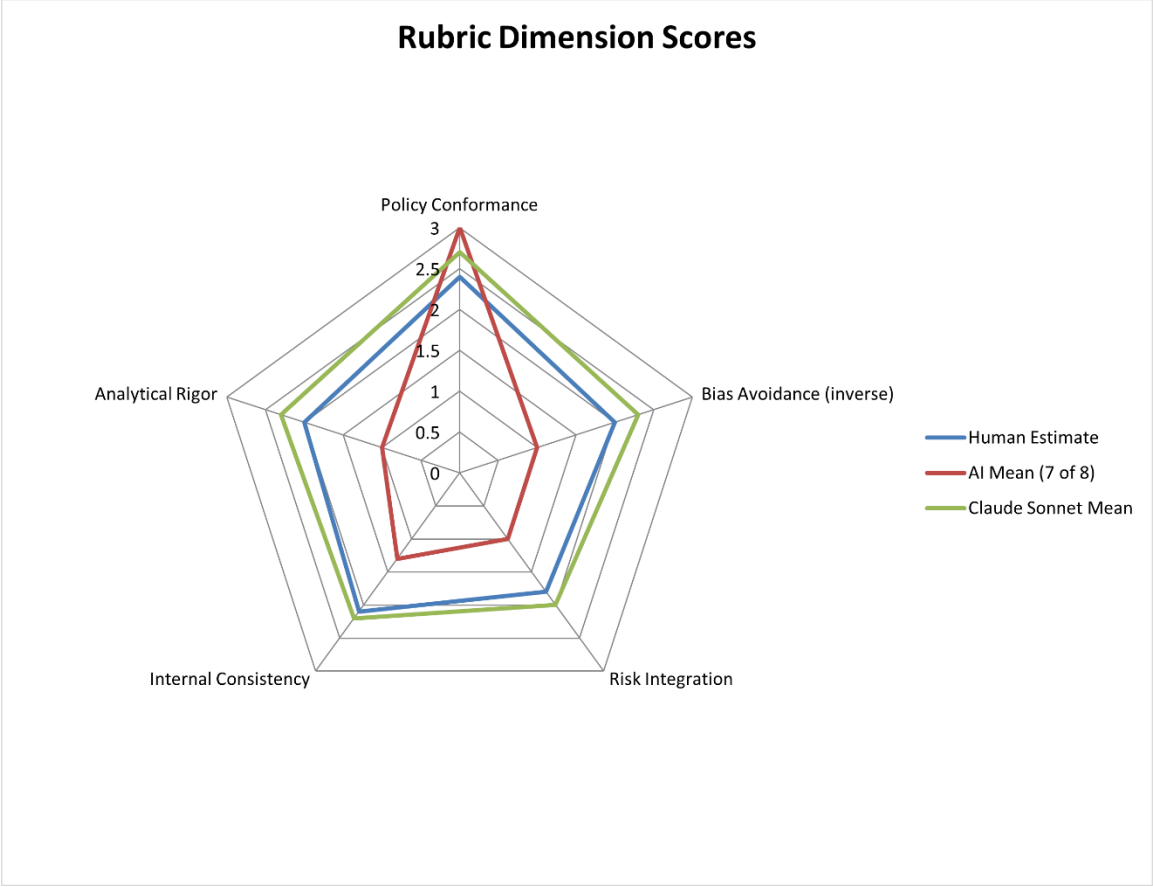

Figure 8. Rubric Dimension Scores Map

Two-proportion z tests take the observations from descriptive to hypothesis-testing. The two-proportion z test compares the percentage of respondents who triggered a bias indicator within one sample to the percentage of respondents who triggered an indicator within another sample. It is used to determine if there is enough evidence to

support that the difference between both samples exceed what chance would allow. Each of the comparisons shown on Table 7 reached significant levels ($p < .001$) with each of the Cohen's h values exceeding 1.00; more than double the common definition of a large effect size as being greater than 0.80. These findings suggest that the differences found in bias indicators between humans and AI are indicative of a significant difference in how humans and AI use competing information to make decisions rather than a minor or uncertain observation.

Table 7. Cognitive Bias Trigger Rates – Human vs. AI Aggregate with z-statistics, p-values, and Cohen's h

Bias Construct	Human Count	Human %	AI Count	AI %	z-statistic	p-value	Cohen's h	Effect
Optimism Bias	14	45.2%	231	96.3%	-9.089	< .001	1.278	Large
Anchoring	13	41.9%	219	91.3%	-7.362	< .001	1.132	Large
Planning Fallacy	12	38.7%	210	87.5%	-6.642	< .001	1.076	Large
Trade-off Difficulty	17	54.8%	231	96.3%	-7.786	< .001	1.084	Large
Confirmation Bias	10	32.3%	210	87.5%	-7.405	< .001	1.211	Large

1. Optimism Bias

When respondents select schedule and cost values that equal or undershoot the APB estimates even when CAIG ICE and Medium-High risk assessment data contradict that choice, this indicator is triggered. 231/240 AI runs (96.3 percent) triggered this bias indicator: seven models chose the 48-month schedule and AUPC in every run (aligning with APB, not CAIG ICE). Claude Sonnet is the anomaly within this dimension. Fifteen of its twenty-one single step runs chose CAIG ICE schedule and cost, and all nine of its incremental runs included CAIG values at some point. Roughly 45 percent of humans triggered this indicator. The majority of human single-step respondents chose the ICE schedule, indicating even responders who chose "full-KPP" realized the original schedule estimate was unrealistic. Seven of eight AI models surpassed the highest subgroup of human respondents in rate of optimism bias indicators.

2. Anchoring

As shown in Figures 5 and 6, the AI models anchored to the APB values: the first complete set of credible numbers the survey case study presented. CAIG ICE values had negligible impact on AI selection behavior. Human respondents presented a more bell-curve distribution, with plurality of single-step respondents adjusting their schedule choice toward the CAIG ICE. Humans adjusted farther from the APB anchor than AI.

3. Planning Fallacy

When respondents select the aggressive schedule (which, if chosen, should trigger awareness of significant schedule risk) while simultaneously rating TRL maturity as important and underweighting CAIG ICE information, this indicator fires. AI chose the 48-month schedule, rated TRL as significantly important, and did not choose CAIG-backed rationales. They took the inside view of program-specific data while ignoring the outside view of analogous historical data offered by CAIG. This replicates the pattern Kahneman and Lovallo (1993) described in which "overly optimistic forecasts result from the adoption of an inside view of the problem, which anchors predictions on plans and scenarios" (p. 17). This entire episode replicates precisely the dynamic Mortlock and Dew (2021) observed when private sector weaponeers proposed the original JCM program back in 2004--2006: PMs become so focused on planning a detailed design that they will build which they know how to build (meaning things are under their control), which in turn makes them optimistic about schedule (and cost) chances that their company cannot actually deliver. The weapons acquisition literature is replete with examples of the planning fallacy. Buehler, Griffin, and Ross's (1997) seminal work on planning fallacy tasks showed that people are institutionally incentivized to focus on their plans more than is statistically rational, which actually increases bias rather than decreases it. This study suggests that AI models may be susceptible to structural parallel: because so much program information exists about the program baseline (CDD, APB, TRL, stakeholder support), but comparatively little exists about analogs/options (a single ICE), AI interprets this pattern as reason to overweight the inside view.

4. Difficulty Making Trade-offs

Single-step implies inability to make trade-offs: when polled, AI models did not demonstrate flexibility to make dotted-line capabilities optional. Faced with a described program that included Med/High integrated system risk, three critical technology elements stuck at TRL 6 each with unique risk profiles, and four threshold platforms with varying degrees of integration complexity, AI chose full-up capability on all counts. In real life, the original JCM program was cancelled six months after Milestone B because the "full steam ahead" strategy was incontrovertibly inexecutable (Mortlock, 2020). Mortlock (2020) identified this precise dynamic as the central challenge of formulating incremental strategies, noting that despite DoD policy preference for incremental development dating back three decades, approval authorities continue to greenlight many programs as single-step development efforts when an incremental approach may be more appropriate. In the JCM case specifically, the warfighter received no capability for over a decade because the Services approved the original single-step strategy and then cancelled it, whereas the follow-on JAGM program adopted an incremental approach that leveraged existing GFE and NDI components to reduce risk, cost, and schedule (Mortlock, 2020).

The follow-on JAGM program achieved success by executing the trade-offs that both JCM program leads and AI models failed to execute. This included decreasing to the dual-mode seeker, utilizing NDI warhead and motor, and reducing platforms in first increment (Mortlock, 2020; Mortlock & Dew, 2021). GAO (2015) found as far back as 2015 that defense acquisition's organizational incentives cause PMs to advocate for strategies that will lead to successful acquisitions (defined as receive approval at milestones) instead of acquisition strategies that are sound (defined as can be executed on cost, schedule, and promised performance parameters and will lead to fielded capability). AI models in this study essentially held a mirror up to the defense acquisition workforce and copied its behavior by selecting the strategy most likely to achieve milestone success instead of selecting the strategy most likely to execute. GAO's (2018) subsequent exploratory statistics study of knowledge-based acquisition practices reinforces this interpretation: programs that capped their development phase at 5–6 years or less AND chose incremental development faced lower CH&S growth than programs that chose unlimited development or chose rapid development over incremental. The AI models in this study selected both.

5. Confirmation Bias

AI models systematically chose rationale options that supported their immediate prior strategy choice while ignoring rationale options contrary to their choice. The survey presented respondents with rationales supporting both the APB's position and the CAIG ICE's position (in random order within the program description body, not isolated as addenda as was done in Phase 2). These were not quotes from human analysts: the survey included them as possible choices. Every respondent had access to the CAIG ICE rationale options, but only humans selected them. This occurred despite the CAIG being an organization charged with creating independent cost estimates (ICEs), a statutory component of the DoD acquisition process. This pattern of using available internal evidence meets the definition of confirmation bias established at the start of this study.

6. Legacy Preference and Recency Bias

AI respondents showed no legacy preference because they never selected existing weapons-system options (COTS) over new development KPP values. The study could not measure recency bias because the design controlled order and recency of information presentation.

F. DIFFERENCES IN ANALYTICAL CHARACTERISTICS

In addition to differences in strategy selection and bias flags, the human and AI data sets diverged along several lines of analysis relevant to the research questions. AI models were orders of magnitude more internally consistent than human survey respondents. Human respondents made strategy selections spread across all 3 options with wide variance in every decision variable. AI model clusters converged on a single strategy with near identical selections across all variables. For strategy choice, performance selections, schedule, and cost, the coefficient of variation was near zero for all runs within a given model. The only exception, Claude Sonnet 4.6, had less within-model variance than the human sample.

1. Decision-Rationale Alignment

AI generated response's selected rationales were consistent to their internal selected strategy. Those AI models which selected an increment strategy based on the APB Schedule

and Cost (increment) selected a rationale that references the APB, POM funding and CDD IOC. The AI was able to maintain consistency by deselecting rationales that contained contradictory evidence as opposed to selecting rationales that bridged cognitive dissonance. Overall, humans were less aligned between decisions/rationales than AI models; however, a sub-set of human respondents who selected incremental strategies demonstrated more sophistication in their reasoning process when they referenced both CAIG ICE and risk assessment as reasons for deferring capabilities.

2. Risk Data Utilization

Humans selected capabilities based on risk data but were inconsistent in this selection. One respondent deferred the warhead due to its medium/high risk rating while another deferred the seeker despite its medium risk rating. While the models included the risk assessment in their decision rationales, they did not differentiate between WBS elements with different risk levels when making trade-off decisions. All models treated TRL 6 as binary indicators of readiness instead of apportioning medium risk (seeker, motor) from medium/ high risk (warhead, integration).

3. Sensitivity to Independent Estimates

By law, the CAIG ICE serves as an independent check on program office optimism. Human decision makers were sensitive to the CAIG ICE to some degree, with the ICE driving schedule and cost selections higher than the APB anchor. AI models exhibited near zero sensitivity to the CAIG ICE if it was at odds with the majority of data points favoring the APB.

4. Response Diversity

The main objective of this analysis is to measure not only which strategies were chosen by each group but also the variety of choices made by each group. Shannon entropy provides a means of capturing both aspects with a single numerical value. Shannon (1948) originally developed entropy as a tool for measuring information in information theory (Shannon, 1948) and has been used extensively in decision making. With respect to an array of three different strategy options, maximum entropy is represented as $H\_max = \log_2(3) = 1.585$ bits. The average entropy generated by humans in their response options is

approximately 97.2 % of the maximum possible entropy (H = 1.54 bits). Conversely, seven out of eight of the AI systems demonstrated no entropy (i.e., zero); and, consequently, no strategic variety. Finally, AI demonstrated only 14.6% (H = 0.23 bits) of maximum entropy. A table summarizing all the data analyzed using entropy can be seen in Table 8. This analysis formalizes the primary conclusion from the study that there is virtually no strategic variation among the responses provided by the various AI systems.

Table 8. Shannon Entropy for Strategy Selection Diversity by Group

Group	Single-Step	Two-Inc	Three-Inc	H (bits)	H / H_max	Interpretation
Human (n=31)	7	13	11	1.5409	97.2%	Near-maximum diversity
Claude Opus 4.6	30	0	0	0.0000	0.0%	Zero (complete uniformity)
Claude Sonnet 4.6	21	9	0	0.8813	55.6%	Moderate diversity
ChatGPT Instant 5.3	30	0	0	0.0000	0.0%	Zero
ChatGPT Thinking 5.4	30	0	0	0.0000	0.0%	Zero
Gemini Thinking	30	0	0	0.0000	0.0%	Zero
Gemini Pro	30	0	0	0.0000	0.0%	Zero
Mistral Small 3 24B	30	0	0	0.0000	0.0%	Zero
Qwen 2.5 32B	30	0	0	0.0000	0.0%	Zero
AI Aggregate	231	9	0	0.2307	14.6%	Low diversity

5. Claude Sonnet as Statistical Outlier

The results of the Fisher's Exact Tests indicate that all primary decision variables for each of the other seven AI models were significantly less than those found with Claude Sonnet; specifically, strategy selection was 70.0 percent vs. 100.0 percent Single-Step ($p < .001$), Schedule Anchoring was 30.0 percent vs. 100.0 percent APB ($p < .001$), and Cost Anchoring was 36.7 percent vs. 100.0 percent APB ($p < .001$).

Two-portion Z-tests, which compare Sonnet to the human values, indicate that there are significant differences between the strategy selected by Sonnet (70.0% Single-Step) and that of humans (22.6%), $z = -3.72$, $p < .001$, Cohen's $h = 0.99$. However, Sonnet's (30% APB) use of APB as a Schedule Anchor is higher than that of humans (56.7% APB), $z = 2.08$, $p = .037$, Cohen's $h = 0.55$. There is also not a significant

difference between Sonnet (36.7% APB) and humans (56.7% APB) when selecting a Cost Anchor as APB ($z = 1.55$, $p = .120$, Cohen's $h = 0.40$). Therefore, Sonnet has demonstrated some ability to be influenced by independent estimates, an ability none of the other AI models have shown, however, Sonnet continues to favor the single-step strategy over humans at almost three times the rate.

Given that Anthropic produces both Sonnet and Opus, but that they differ primarily in their architecture and size, these findings suggest that architectural parameters may influence the decision-making behavior that large language models exhibit relative to acquisition decisions related to LLM's rather than solely based upon the alignment methodology or training data.

G. RUBRIC SCORING

Scores for each group are calculated by applying the five-dimension evaluation rubric to human and AI data. The analysis calculates scores for each group by applying the five-dimension evaluation rubric to human and AI data.

Across Dimension 1 (Policy Conformance), every AI model scored perfectly or nearly perfectly. Seven of eight models chose to defer no KPPs across any runs, checking all applicable threshold KPPs and receiving scores of 3 on policy conformance. Claude Sonnet scored 2 to 3 across its incremental runs because it deferred certain KPPs but accounted for relief. Human respondents who chose the single-step strategy received scores of three on this measure, while incremental respondents received scores of 2 to 3 depending on whether they deferred KPP relief processes. Based on policy conformance alone, AI scored higher than humans.

Across Dimension 2 (Analytical Rigor), scores were flipped. Seven of eight AI models received scores of 1 across all runs: they only selected rationales that cited APB-aligned case study data in their justifications and did not reference or attempt to resolve the discrepancy between the Draft APB cost and schedule estimates and the CAIG ICE values that were both provided to them within the case study data. The survey instrument provided two options that cited ICE values ("Supported by the low CAIG ICE analysis" and "Supported by the high CAIG ICE analysis") to all participants yet were almost

never chosen by the AI models. Claude Sonnet received scores between 2 and 3 because its justifications mentioned the discrepancy between the APB estimates and CAIG estimates provided in the case study. Human respondents displayed greater variance, with single-step respondents who chose ICE values receiving scores of 2 to 3 while incremental respondents who mentioned risk assessments with their CAIG citations received scores of 2 to 3. The human group-level estimate is higher than the AI-7 mean because human rationale selections, as reported in aggregate by Mortlock (2020), more frequently included references to multiple competing data points from within the case study rather than to a single confirming data source.

Across Dimension 3 (Internal Consistency), AI models scored between 1 and 2. Despite AI-derived selections agreeing with each other perfectly (responding the same way every time), the strategy with all threshold KPPs, the 48-month schedule, and $108K cost is not internally consistent with either the CAIG ICE values or Med/High risk rating provided in the case study. It was still possible for AI and human selections to perfectly align with each other while reaching an incorrect, inconsistent conclusion using the case study data. Humans who deferred capability in the incremental strategy while keeping APB cost and schedule consistent scored 3 because they had internal agreement between their selections.

Across Dimension 4 (Risk Integration), seven out of eight AI models scored a 1. Every model ignored the risk labels for individual components and did not use the presented risk data to inform any trade-offs or hard decisions about cost or schedule. This was despite identifying TRL 6 maturity as a generally important consideration. Claude Sonnet received scores of 2 to 3 because it sometimes adjusted either cost or schedule in its incremental responses based on risk labels. Human respondents varied. Roughly half of human respondents scored a 2 or higher by justifying at least one decision with respect to the provided risk information.

Across Dimension 5 (Cognitive Bias Indicators, reverse-scored), seven out of eight AI models scored 1. This means that every time they ran the survey instrument, these AI models triggered at least four of the six available bias indicators (optimism bias, anchoring, planning fallacy, inability to make trade-offs, and confirmation bias were all

triggered by most AI models). Claude Sonnet scored between 2 and 3 across different runs. Human respondents had a larger spread of scores than AI models. Single-step respondents who chose to receive ICE values and mentioned risk information received a score of 2 while incremental respondents with rationale that referenced the provided data scored between 2 and 3.

The analysis calculates scores by summing across each of the five dimensions, presented in Table 9. These scores show that AI models beat humans on policy conformance but were outperformed by humans on analytical rigor, internal consistency, risk integration, and cognitive bias. The combined result was that humans had a higher mean composite score than AI. This effect was mostly driven by scoring gaps on Dimensions 2–5.

Table 9. Rubric Composite Scores

Dimension	Human Estimate (group level)	AI Mean (7 of 8)	Claude Sonnet
Policy Conformance	2.4	3	2.7
Analytical Rigor	2	1	2.3
Internal Consistency	2.1	1.3	2.2
Risk Integration	1.8	1	2
Bias Avoidance (inverse)	2	1	2.3
COMPOSITE TOTAL	**10.3**	**7.3**	**11.5**

Note. Human values are group-level estimates derived from applying the rubric rules to the aggregate frequency distributions reported in Mortlock (2020). Because individual-level human response data was not available, these values do not represent means of 31 individually scored responses and are reported for descriptive comparison only. AI values represent true means of 30 individually scored runs per model.

These dimension-level scores highlight a core paradox illustrated by this research: every AI model achieved perfect scores on policy conformance by selecting every KPP that the CDD called for. However, they achieved this by ignoring almost all of the other information provided to them in the case study. They effectively optimized their responses to select the most policy-compliant answer possible without regard to realism. The most policy-conformant solution was also the solution that ICE flagged as being unrealistic, and the solution that had been previously canceled. The human subgroups that selected incremental strategies scored lower on policy conformance because they

deferred some KPPs. However, they scored higher across the remaining dimensions, demonstrating a willingness to engage with trade-offs among the competing data points rather than optimizing toward a single one. Claude Sonnet's nine incremental runs produced the highest composite scores observed in the entire dataset. This suggests that when an AI model engages with the same trade-off space as the human incremental subgroups, it can equal or exceed human analytical performance on this rubric. Figure 9 shows the composite rubric scores from each of the three groups. These scores serve as a single-number summary of decision quality for the purposes of this study.

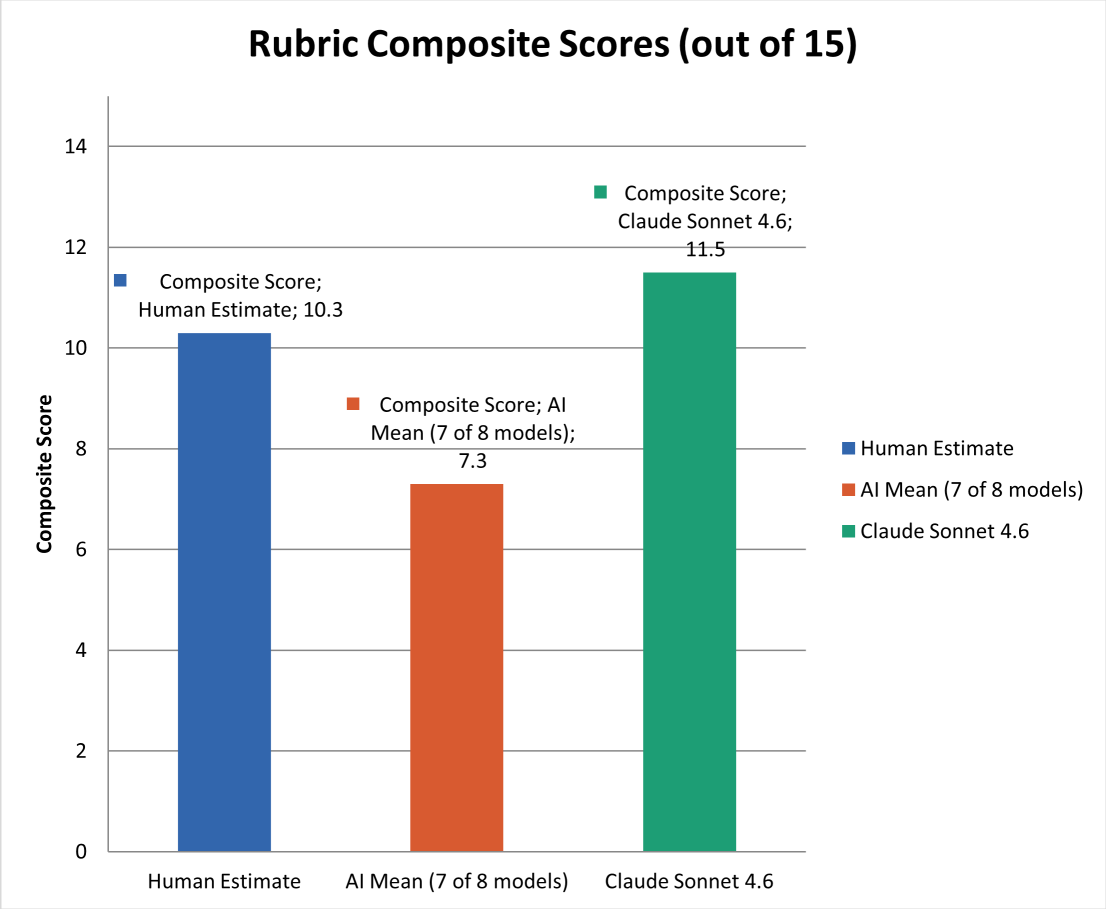

Figure 9. Rubric Composite Scores

Because the respondent human data is aggregated into frequency distributions instead of the actual, individual survey data, the current study could not calculate composite rubric scores at the level of the individual. The score of 10.3 is a group-level estimate based upon how the human respondents selected subgroups in each strategy category. In contrast, the AI composite means (AI-7 M = 7.3 & Claude Sonnet M = 11.5)

were calculated using individually scored run results. Due to the differences between these two approaches to calculating composite rubric means, it was only possible to describe the differences between them; thus, it was not feasible to apply non-parametric statistical procedures (i.e., Mann-Whitney U test) because those types of analyses depend on having the individual-level data for all participants. It is also true however that there is no question about whether or why the difference exists: the mean for AI-7 is significantly lower than the mean for humans, with an absolute difference of 3.0 on a scale ranging from 1 through 15; that difference is due to significantly low scores by AI-7 in three areas of content: Analytical Rigor (1.0 vs. 2.0); Risk Integration (1.0 vs. 1.8); and Bias Avoidance (1.0 vs. 2.0). Future studies that collect individual-level response data from humans will have the opportunity to expand the analytic scope to include composite rubric scores.

H. SUMMARY

This chapter reported on the execution of the comparative case study research. This includes the case study approach, data collection procedures for both human and AI respondents, the rubric development framework, and the results of the comparison. The human baseline data shows that a strong majority of respondents recommended an incremental development approach, with variance across all decision variables. The AI data from 240 runs across eight AI models revealed an almost universal convergence on a Single Step strategy with full KPP performance, the 48-month APB schedule, and APB-aligned cost estimates. Inferential statistical tests confirm that every human-AI comparison on categorical decision variables reaches significance at $p < .001$ with large effect sizes (Cramér's $V = 0.748$ for strategy selection; Cohen's $h > 1.00$ for all five bias constructs). The AI programs reproduced the exact decision-making process as the JCM model that was canceled for its inability to execute. The analysis measured indicators of Optimism Bias, Anchoring, Planning Fallacy, Confirmation Bias, and Difficulty Making Trade-Offs within these models at rates equal to or higher than in humans. The main distinction is that human variance produced risk-mitigating incremental strategies that ultimately proved more viable but also introduced inconsistency. AI consistency, while eliminating noise, also eliminated the trade-off behavior and conservative adjustments

that most human respondents applied. Claude Sonnet 4.6 appears as a statistical outlier among AI models. Sonnet exhibited sensitivity to independent cost estimates that do not differ significantly from human behavior (Fisher's exact $p = .195$) while still diverging significantly from humans in overall strategy formulation (Cohen's $h = 0.99$). Shannon entropy analysis confirms that human respondents achieve 97.2% of maximum strategy diversity. On the other hand, the AI aggregate achieves only 14.6%, with seven of eight models registering zero entropy. Chapter V addresses these findings in its conclusions and recommendations for follow-on research.

THIS PAGE INTENTIONALLY LEFT BLANK

V. FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

Chapter IV demonstrated that AI models, when provided identical inputs to human acquisition professionals, converged almost universally on the same single-step strategy that was approved for the original JCM program and subsequently cancelled, triggering cognitive bias indicators at rates that matched or exceeded those observed in humans. Rather than correcting for the optimism, anchoring, and trade-off avoidance that have plagued acquisition planning for decades, current AI systems reproduced those same patterns through a different mechanism. This chapter compiles those findings into direct responses to the three research questions posed in Chapter I, evaluates the practical implications for integrating AI into acquisition planning, and offers recommendations for both policy and future research.

A. SUMMARY OF FINDINGS

This research compares human outputs against AI-generated outputs using the Joint Common Missile (JCM) case study to determine whether AI can reduce human cognitive bias in the development of ASs and APBs. The study administered the identical survey and case study data that human acquisition professionals participated in (Mortlock, 2020) and eight AI models (240 total runs). This chapter organizes the findings of this research by the original research questions presented in Chapter I.

B. RESPONSES TO RESEARCH QUESTIONS

The following sections address each of the three research questions in the order they were introduced in Chapter I, drawing directly from the statistical results and rubric scoring presented in Chapter IV

1. Primary Research Question

To what extent can AI reduce or mitigate human cognitive bias in the development of Acquisition Strategies and Acquisition Program Baselines?

The results presented here do not support the hypothesis that current AI systems reduce bias in acquisition planning when provided with the same inputs as their human counterparts. Decisions made by the AI models produced optimism bias, anchoring,

planning fallacy, confirmation bias, and inability to make trade-offs decision acceleration trigger at rates that matched or exceeded those observed in the human decisions. The AI models chose overwhelmingly the same single-step, full-KPP, APB-aligned strategy that was approved for the real JCM program in 2004 before being cancelled six months later.

This conclusion requires important qualification. The AI models are biased in the same sense that humans are biased. Human bias is the product of cognitive heuristics, organizational pressure, career incentives, and individual risk preferences. The AI models exhibited functionally similar behavior for a different reason: highly systematic optimization toward the solution supported by most of the data points in the case study when tasked to optimize for a specific outcome without the capability to weight contradictory data (CAIG ICE) appropriately. To the AI models, the CAIG ICE was another piece of data rather than a statutorily mandated independent analysis intended to offset program office optimism.

This distinction has significant implications. Human bias is hard to manage because it is subconscious; as Kahneman (2011) famously showed us, the cognitive heuristics that lead to biased decision-making are driven by intuition-based (fast) thinking that not even expert can easily recognize and avoid. AI weighting tendencies are not hardcoded into the brain, and developers can "fix" them through careful design, prompting, or analytical processes that require the tool to weigh an independent estimate differently than a programmatic estimate. Recent work by Echterhoff et al., (2022) shows that AI can even be designed to recognize anchoring bias across sequential decisions and then flag the biased decision for further review or re-present the information to reduce bias upfront. Acquisition planning could benefit from similar techniques that design AI programs to specifically call out and highlight CAIG-APB differences rather than "solve" them by siding with most data points.

2. Secondary Research Question

How do AI-generated acquisition decisions differ from those produced by human professionals when given the same program data?

The contrasts of decisions are substantial. Human and AI distributions are inverted on strategy selection. Most humans chose incremental strategies while 96.3

percent of AI runs chose single-step strategies. Humans traded off capabilities to mitigate risk while AI models kept every capability in nearly every run. On schedule and cost, humans split their responses between APB and CAIG figures while AI models clung heavily to APB figures. On internal consistency, humans varied wildly across every decision variable while AI models demonstrated near-zero variance across all but one model, Claude Sonnet 4.6.

Together, these patterns point to a central tension: The variance within human responses that created inconsistencies and biases also created the conservative heuristics, risk awareness, and trade-off behaviors that led human respondents to choose viable acquisition strategies. The consistency within AI responses that removed noise also removed the flexibility in decision-making exhibited by most humans. Pure consistency and pure variance do not create ideal acquisition choices; instead, acquisition requires some optimal balance of the two. This intuition aligns with a similar conclusion from Alon-Barkat and Busuioc (2023) that AI should be viewed as neither distrusted nor trusted by human users. The results add to the theoretical framework that Csaszar et al. (2024) developed to understand AI's impact on the strategic decision-making process, specifically the processes of search, representation, and aggregation. Under their framework, AI speeds up and broadens strategic search, but may decrease the diversity and novelty of the strategies considered. The AI models tested here exhibited precisely this behavior, rapidly and consistently converging on the data maximizing strategy while overlooking the wider strategy space that human respondents reached via incremental solutions.

3. Tertiary Research Question

What opportunities and limitations exist for integrating AI tools into the acquisition planning process?

AI systems can quickly generate an analytical baseline. The speed and fidelity with which AI models parsed the case study facts and generated a modeled acquisition strategy recommendation indicate value as a sort of "first-draft" tool that a PM could start analysis with. Secondly the AI models' underweighting of the CAIG ICE was technically correct but yielded the wrong conclusion. Developers could code future AI acquisition

tools with instructions to look for ICEs, appropriately weight them relative to program office estimates, flag variance between program office and independent estimates, and present sensitivity analysis across all authoritative estimates. Finally, analysts can use AI-generated responses as a baseline for explaining their decisions as humans. If we know where the AI came out on a decision point, it can highlight where our professional judgment is moving away from the data-driven analysis. These discussions can facilitate more transparent conversations about which departures are likely productive professional judgments, and which are dangerous cognitive biases.

The AI models did not know a CAIG ICE was not just another "data point." This is likely true for many of the nuances in the acquisition domain suggesting AI general-purpose models need significant adaptation for use in acquisition. AI models failed to make trade-offs. This is perhaps the most critical shortfall of this exercise. Making trade-offs is likely the single most valuable expertise a PM contributes to acquisition strategy development. The AI models provided no evidence of this ability and may be inherently incapable of providing guidance beyond what is explicitly spelled out in the information it is provided. All the AI model answers converged near perfectly on the same strategy. This sort of conformity could easily lead decision-makers into automation bias if organizations use these tools for decision support. Alon-Barkat and Busuioc (2023) did not find evidence that bureaucratic officials would blindly defer to algorithmic decisions. However, they did find evidence that officials were selectively more likely to comply with AI recommendations that aligned with their beliefs. In acquisition, a PM who believes the program can only be executed successfully with all KPPs requiring a single step may be concerned that breaking the KPPs into phases would prolong schedule but might be unattached from that skepticism if the AI sends the same recommendation. Miedema et al. (2026) found that subjects who received AI-generated explanations were significantly more likely to trust the decision as correct, even when the decision was incorrect. This makes the quality of AI reasoning in addition to the conclusions it draws important to the responsible use of these tools.

C. POTENTIAL ROLE OF AI IN ACQUISITION PLANNING

The conclusion from this research is that current AI systems are not ready to be standalone planners for selecting acquisition strategies but could be powerful tools for developing strategies if carefully tailored and bounded. When allowed to choose, AI systems lack the necessary situational awareness to select the correct decision. On the other hand, when forced into a structured output referencing underlying data, AI systems can provide decision-value when combined with a judgment that recognizes where the analysis falls short or is misaligned.

AI systems making the same decision as the original JCM program, rather than the corrective decision that human survey respondents reached, shows that AI systems do what they are told. If you provide data to an AI system that reinforces program office inputs more strongly than it does independent estimates, the AI system will output a solution that leans towards the program office. This is not a failure of AI, but a reflection of the informational environment in which acquisition professionals make decisions. Between the CDD, APB, POM, and signature collection of stakeholders, the paper trail of an acquisition program is one heavily weighted towards the program office viewpoint. The CAIG ICE is an essential voice of dissent, but it is only one voice. Input-heavy AI systems will inherently de-weight the independent estimate.

Perhaps the best use of AI, then, is not as an optimizer but instead as a decision-forcing tool that clearly articulates the discord in available data. An AI that can highlight where the program office and independent estimates diverge, provide alternative courses of action at different risk tolerances, show sensitivity to key variables, and offer trade space across capability, cost, and schedule could help correct the very issue that caused both the surveyed humans and the AI system to make poor decisions.

D. RECOMMENDATIONS FOR POLICY AND PRACTICE

The following recommendations translate the study's findings into potential actionable guidance for how the acquisition community can integrate AI tools without replicating the bias patterns this research identified.

1. AI as Decision Support, Not Decision Maker

This research further supports the argument that decisionmakers should use AI as an assistant, not as a decisionmaker itself. While the inherently structured, referenced-from-data output of the tool made it easy to see where AI models added value alongside human judgment, their consistent convergence toward the acquisition strategy that has repeatedly failed in practice showed that these tools are not yet ready to make acquisition decisions on their own.

2. Structured Analytical Frameworks for AI-Assisted Planning

Developers should design acquisition planning support AI tools using structured analytical techniques that force the model to do analysis (rather than simply make a judgement call). Examples include analysis to highlight program office / independent estimate variance both quantitatively and qualitatively, development of alternative acquisition strategies across a relevant spectrum of risk appetite levels, sensitivity analysis demonstrating how responsive the chosen strategy is to variation in key assumptions, and trade-off matrices that visually display the cost, schedule and capability impacts resulting from deferring known technologies / platforms to follow-on increments.

3. Independent Estimation Integration

It is worth mentioning that the consistent discounting of CAIG ICE by the AI models is as much a symptom of a larger problem. Policy makers should consider requiring acquisition planning tools (AI-augmented or otherwise) to both explicitly highlight program / independent estimate variance and to show analysis across the full spectrum of estimates provided rather than choosing one (e.g. CAIG ICE) point estimate to design to. The independent estimate exists because history has shown time and again that program offices demonstrate a consistent pattern of optimistic estimations.

4. Lessons Learned Supporting the Study of Behavioral Acquisition

Finally, this research supports Mortlock's (2020) call for a new field of study: behavioral acquisition. Mortlock (2020) proposed behavioral acquisition as a research field that combines the study of program management, organizational dynamics, defense

acquisition, and psychology within acquisition decision making. This is analogous to behavioral finance, which studies both economics and psychology within financial decision making. This research addresses this call by providing the first empirical comparison of human and AI acquisition decisions using the same inputs. The results indicate that AI Models can demonstrate similarly biased indicators but via differing structures than those demonstrated by humans.

Additionally, Mortlock (2020) recommended three specific policy changes. These changes include making incremental development the default strategy requiring MDA justification for single-step approaches, augmenting component TRLs with risk ratings for milestone decisions, and specifically addressing integration risk at all milestone reviews. The AI models' failure to make trade-offs in this study reinforces the need for the first recommendation, as both AI and human decision makers demonstrated reluctance to deviate from the full-capability, single-step approach without explicit policy mechanisms compelling them to consider alternatives. As Mortlock and Dew (2021) describe, Mortlock previously demonstrated through analysis of the JCM, Enhanced Combat Helmet, and Army infantry combat vehicle programs that planning fallacy, over-optimism, inability to make trade-offs, and recency bias are pervasive problems in defense acquisition. The discovery here that AI can produce results with functionally similar indications of bias through different mechanisms warrant expanding the behavioral acquisition field to include study of the behaviors of AI models commonly used to support acquisition decisions. Similar to how Mortlock and Dew (2021) showed that factors at the institutional, Service, and program levels amplify or attenuate human cognitive biases, researchers need to investigate how the weighting patterns that AI models exhibit intersect with human cognitive biases. Kiesling and Chong's (2020) discovery that the language GAO uses when writing reports about acquisition programs is consistent with cognitive bias effects on human behavior can be used as a starting point for this type of analysis on AI-generated acquisition documentation.

E. RECOMMENDATIONS FOR FUTURE RESEARCH

Future researchers can extend this study in several ways. First, repeat this analysis with additional cases spanning several acquisition categories, acquisition pathways, and

program types to assess if patterns continue throughout the defense acquisition portfolio. Second, explore hybrid human-AI decision making in which researchers provide the AI-created analysis to the human decision maker along with the case study, and evaluate performance to see if hybrid teams make better decisions by leveraging AI reliability and human judgment. Fourth, run the sensitivity analysis to include persona, prompt tone, and temperature to see if these levers influence AI acquisition decisions in a manner that can produce the missing trade-off exploration that was absent in baseline prompt responses. Fifth, use adversarial prompting with AI agents for acquisition planning, in which one AI creates a proposed plan and a second AI that interrogates this plan with independent cost estimates, past failures, and risk information Sixth, research how humans interacting with AI systems for acquisition planning react to the system's proposed acquisition strategy, testing for automation bias and identifying the conditions under which humans accept, modify, or reject AI recommendations.

F. BROADER IMPLICATIONS FOR THE DOD ACQUISITION DECISION-MAKING PROCESS

The defense acquisition system spends hundreds of billions making decisions in the early planning stages of programs. The DoD currently has planned investments totaling nearly $2.4 trillion across its highest cost weapon programs (GAO, 2025). Research going back decades has consistently found cost growth, schedule delays, and performance shortfalls result from poor planning and early decision making on programs (GAO, 2025; Drezner & Krop, 1997; Wong et al., 2022). Drezner and Krop's (1997) research showed once baselines are established, they drive funding through the PPBE process and set expectations that become exponentially more costly to adjust. Incremental improvements to the quality of those initial decisions can have outsized impacts. The results from this research show that improvement doesn't come from asking whether to use human judgment or AI, but how to build systems that capitalize on each of their strengths.

The human respondents in this research exhibited a behavior that the AI models could not: a willingness to reject the answer that best fit the data if their professional judgment told them that data supported answer was incorrect. Human respondents were

presented with a program where every technology was at TRL 6, there was 100% stakeholder support, funding had been approved, and the JROC had approved the program's CDD. And they correctly decided that the single-step strategy was too risky. The AI models received the exact same data and decided that the single-step strategy was the best option. But they were wrong for the same reason the original JCM decision-makers were wrong in 2004. Figure 8 and Table 8 put that number into context by showing that 96.3% of AI runs chose the same strategy, Table 10, that was ultimately cancelled in 2004. In contrast, 77.4% of human respondents chose a decision in the incremental approach category that matched what JAGM ultimately did in 2015.

Table 10. Historical Strategy Alignment by Decision Group

Decision Group	Aligned with JCM (Cancelled) %	Aligned with JAGM direction (Incremental) %
AI Models (240 runs)	96.3%	3.7%
Human Respondents	22.6%	77.4%

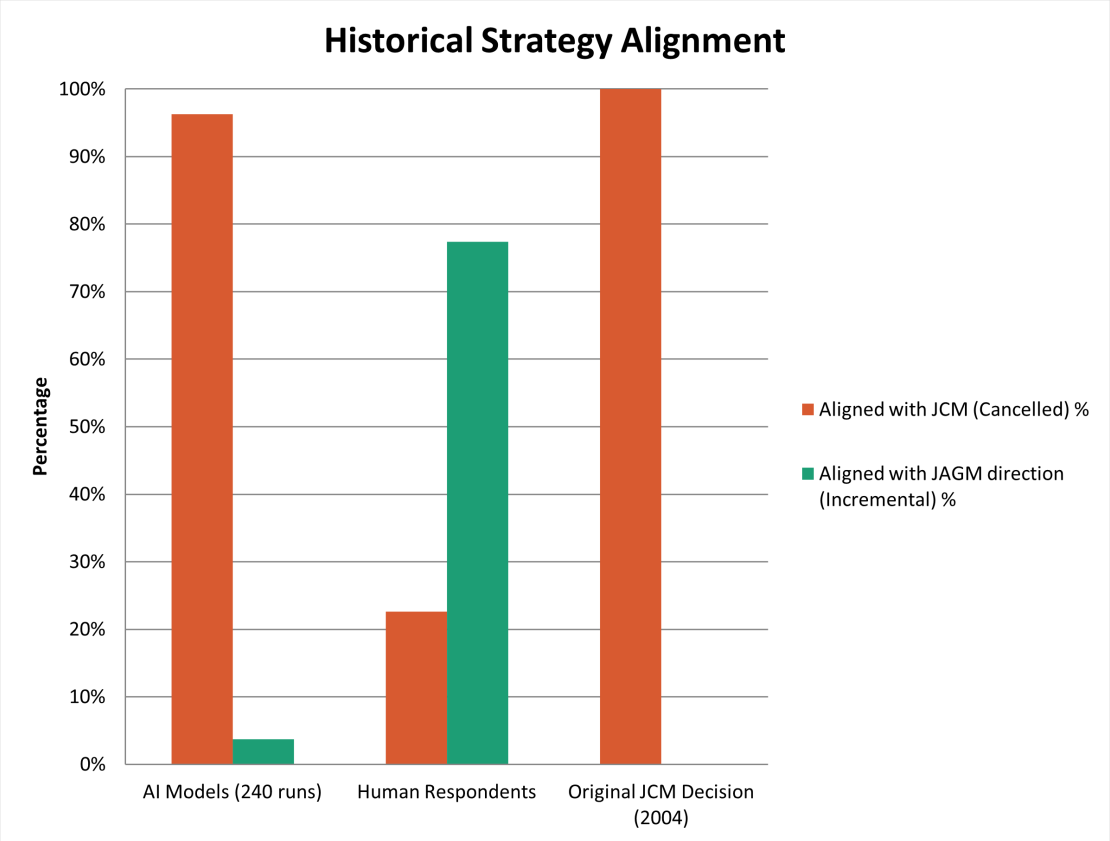

Figure 10. Historical Strategy Alignment

Table 11. Historical Strategy Alignment – JCM vs. JAGM

Strategy	JCM (2004) — CANCELLED	JAGM (2015) — SUCCESSFUL
Approach	Single Step	Incremental (2 increments)
Seeker	Tri-mode (full KPP)	Dual-mode (Inc I), Tri-mode (Inc II)
Warhead	Multipurpose (full KPP)	Hellfire NDI (Inc I), Multipurpose (Inc II)
Motor	Common (full KPP)	Hellfire NDI (Inc I), Common (Inc II)
Platforms	AH-64, AH-1Z, MH-60, F/A-18	AH-64, AH-1Z only (Inc I)
EMD Schedule	48 months	24 months (Inc I)
AUPC	$108K-$120K (APB)	Reduced scope, reduced cost
Outcome	Cancelled 6 months after MS B	Successfully fielded

This is not to say human judgment is superior to AI. The human respondents showed their own form of significant bias: variance that cannot be explained by the data, inconsistent use of risk data when making decisions, and struggled to prioritize between competing constraints. Rather, human judgment and AI have significant blind spots, but they are different blind spots. When AI-human decision making is brought together in a system where each type of failure is visible to the other, there is great potential to improve acquisition outcomes.

There is clearly a need for additional research in this area. As AI is introduced across the DoD it will be critical that AI-assisted decision-making tools are being built to avoid replicating human thought patterns that have led to decades of cost growth, schedule delays, and shortfalls in capability. Defense acquisition has remained on the GAO's high-risk list since they first established it in 1990 (GAO, 2025). In their analysis of persistent problems in acquisition, Mortlock and Dew (2021) concluded that underappreciated and understudied are the cognitive biases that play into the root causes of acquisition program failures. Cited examples include the planning fallacy, inability to make effective trades, over-optimism, and recency bias. This research is the first step in showing empirical evidence that AI systems, when provided the same information as their human acquisition counterparts, will reproduce systemic behavioral biases rather than course correct for them. The more we understand how these biases affect both human and AI decision making, the better we can design AI to support human decision-

making that mitigates risk of program failure and gets capabilities into the hands of the warfighter faster and more efficiently.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Alon-Barkat, S., & Busuioc, M. (2023). Human-AI interactions in public sector decision making: Automation bias and selective adherence to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169.

Ayvaz, S., & Alpay, K. (2021). Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. *Expert Systems with Applications, 173*, 114598. https://doi.org/10.1016/j.eswa.2021.114598

Buehler, R., Griffin, D., & Ross, M. (1997). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In M. Bazerman, D. Messick, A. Tenbrunsel, & K. Wade-Benzoni (Eds.), *Environment, ethics, and behavior* (pp. 250–270). New Lexington Press.

Clinger-Cohen Act of 1996, Pub. L. No. 104–106, Division D, 110 Stat. 186. https://www.congress.gov/bill/104th-congress/senate-bill/1124

Csaszar, F. A., Ketkar, H., & Kim, H. (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Science, 9*(4), 322–345. https://doi.org/10.1287/stsc.2024.0190

Defense Acquisition Workforce Improvement Act of 1990, Pub. L. No. 101–510, Title XII, 104 Stat. 1485. https://www.congress.gov/bill/101st-congress/house-bill/4739

Department of Defense. (2020). *Operation of the Adaptive Acquisition Framework* (DoDI 5000.02). Under Secretary of Defense for Acquisition and Sustainment. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.PDF

Department of Defense. (2021). *Major capability acquisition* (DoD Instruction 5000.85, incorporating change 1, November 4, 2021). Office of the Under Secretary of Defense for Acquisition and Sustainment.

Department of Defense. (2022a). *The defense acquisition system* (DoD Directive 5000.01, incorporating change 1, July 28, 2022). Office of the Secretary of Defense.

Department of Defense. (2022b). *Operation of the adaptive acquisition framework* (DoD Instruction 5000.02, incorporating change 1, June 8, 2022). Office of the Under Secretary of Defense for Acquisition and Sustainment.

Department of Defense. (2024). *Cost analysis guidance and procedures* (DoD Instruction 5000.73). Office of the Under Secretary of Defense (Comptroller).

Department of Defense Office of Inspector General. (2025). *Summary report: Lessons learned from DoD OIG reports on acquisition oversight* (Report No. DODIG-2025-155).

Drezner, J. A., & Krop, R. (1997). *The use of baselining in acquisition program management*. RAND Corporation.

Echterhoff, J. M., Yarmand, M., & McAuley, J. (2022). AI-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In CHI *Conference on Human Factors in Computing Systems* (CHI '22), April 29–May 5, 2022, New Orleans, LA, USA. ACM. https://doi.org/10.1145/3491102.3517443

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates.

Federal Acquisition Streamlining Act of 1994, Pub. L. No. 103–355, 108 Stat. 3243. https://www.congress.gov/bill/103rd-congress/senate-bill/1587

Fox, J. R. (2011). *Defense acquisition reform, 1960–2009: An elusive goal*. Center of Military History.

Flyvbjerg, B., Garbuio, M., & Lovallo, D. (2009). Delusion and deception in large infrastructure projects: Two models for explaining and preventing executive disaster. *California Management Review, 51*(2), 170–194.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Government Accountability Office. (2015). *Defense acquisitions: Joint action needed by DoD and Congress to improve outcomes* (GAO-16-187T).

Government Accountability Office. (2018). *Weapon systems annual assessment: Knowledge gaps pose risks to sustaining recent positive trends* (GAO-18-360SP).

Government Accountability Office. (2020). *Cost estimating and assessment guide: Best practices for developing and managing program costs* (GAO-20-195G).

Government Accountability Office. (2025). W*eapon systems annual assessment 2025: DoD leaders should ensure that newer programs are structured for speed and innovation* (GAO-25-107569).

Hammond, J. S., Keeney, R. L., & Raiffa, H. (2006). The hidden traps in decision making. *Harvard Business Review, 84*(1), 118–126.

Harris, C. G. (2020). Mitigating cognitive biases in machine learning algorithms for decision making. In *Companion Proceedings of the Web Conference 2020* (pp. 775–781). ACM.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39*(1), 17–31.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.

Kiesling, J. K., & Chong, D. M. (2020). *Examining the role of decision biases in shaping acquisition decisions within defense acquisition programs* [Master's thesis, Naval Postgraduate School]. Calhoun: The NPS Institutional Archive. https://calhoun.nps.edu/entities/publication/9932e198-ab76-4152-af69-63f730b153e6

Koszykowski, M., & Orzeszko, W. (2025). Machine learning in project schedule creation: A systematic literature review. *Journal of Scheduling.* https://doi.org/10.1007/s10951-025-00857-w

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lau, H. C. W., Choy, K. L., Lau, P. K. H., Tsui, W. T., & Choy, L. C. (2004). An intelligent logistics support system for enhancing the airfreight forwarding business. *Expert Systems, 21*(5), 253–268.

Miedema, E., Waschull, S., & Emmanouilidis, C. (2026). Towards trustworthy artificial intelligence for decision-making: A life cycle perspective on knowledge- and data-driven artificial intelligence systems. *Computers in Industry, 174*, 104409. https://doi.org/10.1016/j.compind.2025.104409

Mini, T. V. (2026). Understanding machine learning: Real-world examples that make sense. *International Journal of Information Technology Research Studies (IJITRS)*, *2*(1), 1–12. https://doi.org/10.5281/zenodo.18479380

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.

Mortlock, R. F. (2020). Studying acquisition strategy formulation of incremental development approaches. *Defense ARJ*, *27*(3), 264–311. https://doi.org/10.22594/dau.19-845.27.03

Mortlock, R. F., & Dew, N. (2021). Behavioral biases within defense acquisition. In *Proceedings of the Eighteenth Annual Acquisition Research Symposium* (pp. 93–114). Naval Postgraduate School.

Narbaev, T., Hazir, Ö., Khamitova, B., & Talgat, S. (2024). A machine learning study to improve the reliability of project cost estimates. *International Journal of Production Research, 62*(12), 4372–4388. https://doi.org/10.1080/00207543.2023.2262051

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simon, H. A. (1957). *Models of man: Social and rational*. Wiley.

Stebbins, D., Girven, R. S., Parker, T., Deen, T., De Bruhl, B., Ryseff, J., Welburn Paige, J., Yu Kleiman, A., Bhatt, S. D., Sousa, É. M., Kepe, M., & Fay, M. (2024). *Exploring artificial intelligence use to mitigate potential human bias within U.S. Army intelligence preparation of the battlefield processes* (RRA2763-1). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2763-1.html

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Weapon Systems Acquisition Reform Act of 2009, Pub. L. No. 111–23, 123 Stat. 1704 (2009). https://www.congress.gov/bill/111th-congress/senate-bill/454

Wong, J. P., Younossi, O., LaCoste, C. K., Anton, P. S., Vick, A. J., Weichenberg, G., & Whitmore, T. C. (2022). *Improving defense acquisition: Insights from three decades of RAND research*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1670-1.html

Acquisition Research Program
Department of Acquisition, Finance and Manpower
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net