EXHIBIT B-18 Project Canary AI Evaluation at the Intelligence-Philanthropy Nexus

Project Canary: AI Evaluation at the Intelligence-Philanthropy Nexus

A $38 million collaboration between METR and RAND Corporation reveals the convergence of AI safety evaluation, defense policy, and effective altruist philanthropy—with former IARPA Director Jason Matheny as the central connecting figure.

Project Canary represents a formalization of the relationship between the AI safety evaluation community and the national security establishment. Announced October 9, 2024, the partnership brings together METR—a nonprofit AI evaluation organization with roots in the effective altruism movement—and RAND Corporation, a storied defense think tank with $435.8 million in annual revenue (75% from U.S. government sources) and deep intelligence community ties through its four Federally Funded Research and Development Centers. The collaboration is funded entirely through philanthropy, yet creates direct channels for evaluation frameworks to inform defense and intelligence decision-making. Jason Matheny, RAND's CEO and former Director of the Intelligence Advanced Research Projects Activity, serves as the bridge between these worlds—receiving discretionary funding from effective altruist philanthropy while leading an organization conducting hundreds of millions in classified intelligence work.

The fundamental misconception: Not a joint partnership, but philanthropic funding

The widely circulated description of Project Canary contains a critical factual error. The $38 million does not represent "$17M from METR, $21M from RAND" as bilateral contributions. Rather, **The Audacious Project**—a collaborative philanthropic funding initiative housed at TED—catalyzed approximately $38 million in external funding for the RAND-METR collaboration, with $17 million specifically allocated to support work at METR. The funders include the Bill & Melinda Gates Foundation, MacKenzie Scott, ELMA Philanthropies, Emerson Collective, Skoll Foundation, and Valhalla Foundation. Additionally, Open Philanthropy—the effective altruist organization funded by Facebook co-founder Dustin Moskovitz—committed a separate $10 million over three years to RAND specifically for the Canary collaboration.

This funding structure has significant implications. Project Canary operates as a philanthropically-funded research collaboration, not a government contract or bilateral institutional partnership. METR maintains its claimed independence from AI company funding while entering partnership with an organization conducting extensive classified intelligence work. The philanthropic intermediary—The Audacious Project and Open Philanthropy—provides a layer of separation from direct government funding while enabling coordination between AI safety evaluation and defense policy communities.

Project objectives and the path to mandatory evaluation by 2027

Project Canary's stated mission is developing tools that all frontier AI developers can use to evaluate their systems for risks prior to release. The collaboration targets three focus areas through 2027: developing advanced AI evaluation methods, testing systems for dangerous capabilities, and sharing findings to inform evidence-based policies. The explicit goal is enabling any new AI system to undergo rigorous safety evaluations before release by 2027—effectively establishing evaluation as standard practice.

METR's role focuses on assessing frontier AI systems' ability to act as autonomous agents, particularly evaluating autonomous capabilities that could accelerate AI research and development itself. RAND contributes expertise in assessing potential for AI system misuse, bringing its 75-year history of defense and security research to bear on evaluation methodology. The organizations will share methodologies openly with the wider AI research community and share findings with "decisionmakers"—a term encompassing AI developers, government entities, policymakers, and companies scaling AI systems.

The 2027 timeline is significant. It positions Project Canary to establish evaluation infrastructure just as regulatory requirements crystallize. President Biden's Executive Order 14110 on AI (October 2023) requires red-team testing of dual-use foundation models. The Bureau of Industry and Security is developing reporting requirements. International coordination through the AI Safety Institute Network spans the UK, US, Japan,

France, Germany, Italy, Singapore, South Korea, Australia, Canada, and EU. METR has already submitted formal comments on proposed regulations. Project Canary appears designed to provide the technical capacity and policy frameworks that emerging regulations will require—positioning RAND-METR as the de facto standard for AI safety evaluation.

Jason Matheny: The nexus between intelligence agencies and AI safety

Jason Gaverick Matheny is the central figure connecting Project Canary to the intelligence community. His career trajectory spans academic research on existential risks, leadership of intelligence agencies, White House technology policy, and now command of a major defense think tank—all while maintaining connections to the AI safety and effective altruism communities.

Intelligence community leadership. Matheny joined the Intelligence Advanced Research Projects Activity in 2009 as Program Manager, creating pioneering forecasting programs including the Aggregative Contingent Estimation (ACE) Program and Open Source Indicators (OSI) Program. He progressed to Director of IARPA in 2015, serving until 2018 while simultaneously holding the title Assistant Director of National Intelligence. At IARPA, he led high-risk, high-payoff research for the U.S. intelligence community, overseeing programs in quantum computing, superconducting computing, machine learning, and anticipatory intelligence. He received the Intelligence Community's Award for Individual Achievement in Science and Technology and the National Intelligence Superior Service Medal.

White House technology and national security policy. After leaving IARPA, Matheny founded the Center for Security and Emerging Technology at Georgetown University in 2018/2019, building it into a major player in emerging technology policy research. Simultaneously, he served as a Commissioner on the National Security Commission on Artificial Intelligence from 2018-2021, appointed by Congress and led by Eric Schmidt. The Commission's final report warned that the U.S. government was not prepared to defend or compete against China in the AI era and called for accelerated AI adoption in the Department of Defense by 2025.

In March 2021, Matheny joined the Biden administration in a "triple-hatted" position: Deputy Assistant to the President for Technology and National Security (National Security Council), Deputy Director for National Security (Office of Science and Technology Policy), and Coordinator for Technology and National Security (NSC). In these roles, he led White House policy on technology and national security, bridging NSC and OSTP portfolios on AI, semiconductors, synthetic biology, and U.S.-China technology competition.

RAND Corporation leadership. In June 2022, Matheny was selected from nearly 200 candidates to become the sixth President and CEO of RAND Corporation, effective July 5, 2022. He now leads a 2,000-person research organization with over $435 million in annual revenue and extensive defense and intelligence contracts. Under Matheny's leadership, RAND has dramatically expanded AI research and received over $31.4 million from Open Philanthropy, including $10.5 million in discretionary funding for emerging technology initiatives and $10 million specifically for Project Canary.

Direct connection to Paul Christiano and METR. Matheny and Paul Christiano—founder of the Alignment Research Center, which incubated METR—served together as two of five initial trustees of Anthropic's Long-Term Benefit Trust in 2023. This independent governance body holds authority to elect and remove Anthropic board members, designed to ensure the company pursues AI development for long-term benefit of humanity. Matheny stepped down in December 2023 to avoid conflicts of interest with RAND's policy initiatives. Christiano stepped down in April 2024 to become Head of AI Safety at the U.S. AI Safety Institute (now Center for AI Standards and Innovation). This shared governance role demonstrates direct working relationships spanning AI labs, AI safety research, and now government positions.

Publications on AI evaluation and regulation. Matheny has testified before Congress multiple times on AI challenges and opportunities. In March 2023 testimony to the Senate Homeland Security Committee, he warned: "The U.S. government is not prepared to defend the United States in the coming AI era" and expressed concern about "AI being applied to the development of new cyber weapons and bioweapons." In June 2023, he told the House Science Committee that "AI systems have security and safety vulnerabilities and a major AI-related accident in the United States or misuse could dissolve or lead, much like nuclear accidents set back the acceptance of nuclear power."

In an August 2023 Washington Post op-ed, Matheny proposed a three-part regulatory framework: licensing large concentrations of AI chips, oversight of training data and model capabilities during training, and rigorous review by regulators or third-party evaluators before model release. This framework aligns closely with Project Canary's objectives—developing third-party evaluation capacity that could satisfy regulatory requirements. His academic background includes a seminal 2007 paper "Reducing the Risk of Human Extinction" and pioneering research on cultured meat, demonstrating long-standing focus on catastrophic and existential risks.

RAND's $832.5 million in intelligence contracts and classified research

RAND Corporation operates four Federally Funded Research and Development Centers with extensive defense and intelligence relationships. In FY2024, RAND's revenue totaled $435.8 million, with approximately $328 million (75%) from U.S. government sources. Three major verified contracts total $832.5 million over their respective periods, though annual obligations are lower given the multi-year timelines.

The $417 million DOD intelligence policy research contract. In February 2021, RAND's National Defense Research Institute received a five-year, $417 million indefinite-delivery/indefinite-quantity contract from Washington Headquarters Services for policy research and analysis supporting the Department of Defense's acquisition and sustainment office. The contract covers acquisition and technology, forces and resources, cyber and intelligence, international security and defense, and U.S. Navy and Marine forces. It assists decision-making at the Office of the Secretary of Defense, Joint Chiefs of Staff, Department of the Navy, and unified combatant commands. The contract continues through December 31, 2025.

The $184 million Army research contract. In September 2021, RAND's Arroyo Center—the U.S. Army's sole FFRDC for studies and analysis—received a nine-year, $184.2 million cost-plus-fixed-fee contract from U.S. Army Contracting Command for research and analysis work. The contract extends through September 20, 2030, with funds and work locations determined on individual orders.

The $231 million Air Force contract. In March 2016, RAND's Project Air Force received a five-year, $231.3 million indefinite-delivery/indefinite-quantity contract covering advisory and assistance services. Research areas included strategy and doctrine, force modernization and employment, manpower and training, resource management, cybersecurity, space warfare, and air power requirements. This contract ran through March 2021; renewal contracts are likely but not publicly documented.

Classified research programs and intelligence community clients. RAND has a 75-year history of classified intelligence work dating to its founding in 1946 as "Project RAND" for the Air Force. The organization pioneered systems analysis, nuclear deterrence theory (including "mutually assured destruction"), early satellite and space-based intelligence, and foundational computing research. RAND's National Security Research Division explicitly supports the Office of the Director of National Intelligence, various Intelligence Community agencies, National Geospatial-Intelligence Agency, Defense Counterintelligence and Security Agency, and IARPA itself.

The organization's FFRDC status provides "in-the-family" access to classified information and decision-makers. Specific current classified programs are not publicly disclosed, but RAND's FFRDCs inherently conduct classified research in cyber and intelligence operations, special operations support, nuclear weapons policy, intelligence collection methodologies, counterterrorism strategy, and advanced weapons systems analysis. Much of this work never appears in public contract databases or annual reports.

AI-related research and the Open Philanthropy controversy. RAND has received over $15 million from Open Philanthropy for AI-related research since Matheny became CEO. Major grants include $10.5 million for emerging technology initiatives, $6 million for the Technology and Security Policy Center, and $10 million for Project Canary. This funding has generated controversy. In December 2023, the House Science Committee sent a bipartisan letter to NIST raising concerns about RAND's involvement in drafting President Biden's AI Executive Order, noting "research that has failed to go through robust review processes, such as academic peer review." In September 2024, Senator Ted Cruz questioned RAND's "outsize role in drafting the AI Executive Order." Internal RAND employees reportedly expressed concerns about objectivity due to Open Philanthropy ties.

The financial architecture: Open Philanthropy as central funding node

While Project Canary received no identified direct government funding, Open Philanthropy serves as a central financial node connecting effective altruist AI safety research with defense policy institutions. The organization, founded by Facebook co-founder Dustin Moskovitz and Cari Tuna, describes itself as the leading funder of AI alignment research and has directed over $4 billion in grants as of June 2025.

Open Philanthropy's extensive RAND funding. Since Matheny became RAND CEO in July 2022, Open Philanthropy has committed at least $31.4 million to RAND:

- $10 million over 3 years specifically for Project Canary
- $10.5 million for emerging technology initiatives (including China studies center, Pardee RAND Graduate School support, emerging tech research fund)
- $5.5 million for technology policy training program and research fund
- $3.4 million over 2 years for forecasting platform development
- $2.5 million for general discretionary support

Significantly, much of this funding is at Matheny's discretion, suggesting personal relationships and trust between Open Philanthropy leadership and Matheny. Open Philanthropy's Luke Muehlhauser leads AI governance grantmaking and recommended the RAND grants.

METR's funding independence claim. METR emphasizes financial independence from AI companies: "METR has not accepted funding from AI companies, though we make use of significant free compute credits." The organization states: "Independent funding has been crucial for our ability to pursue the most promising research directions and set standards for evidence-based understanding of risks from AI." Beyond The Audacious Project's $17 million, METR received $220,000 from Longview Philanthropy in 2023 and was initially incubated at Paul Christiano's Alignment Research Center, which received $265,000 from Open Philanthropy in March 2022.

Ajeya Cotra as bridge figure. Ajeya Cotra, a researcher at Open Philanthropy working on AI timelines and AI governance, serves as a METR advisor. This creates a direct connection between Open Philanthropy's funding decisions and METR's strategic direction. The advisory relationship suggests alignment between Open Philanthropy priorities and METR's work—not surprising given that both organizations emerged from the effective altruism movement's focus on existential risk reduction.

The effective altruism ecosystem. Project Canary sits at the intersection of effective altruist funding and traditional defense establishment institutions. METR founder Beth Barnes came from OpenAI and DeepMind alignment research embedded in EA culture. Paul Christiano is a central figure in EA AI safety. Longview Philanthropy is EA-aligned. Open Philanthropy is the EA movement's largest funder. The Audacious Project, while more mainstream, selected Project Canary from among proposals emphasizing global catastrophic risk reduction. RAND, by contrast, represents the defense and intelligence policy establishment with 75 years of Pentagon and CIA relationships. Project Canary formalizes the bridge between these previously separate worlds.

METR's limited AI lab partnerships and independence claims

METR's relationships with major AI labs are far more limited and informal than often assumed in public discourse. The organization conducted pre-deployment evaluations of GPT-4 in early 2023 and Claude 2 in mid-2023, but these were explicitly not formal audits. In 2024, METR published a clarification stating: "METR has not intended to claim to have audited anything" and emphasized that past work consisted of "informal pilots" or "research collaborations" under NDA with no guaranteed influence on deployment decisions.

The OpenAI relationship. METR (then ARC Evals) evaluated GPT-4 in early 2023, receiving early access to multiple versions but could not fine-tune the model and did not have access to the final deployed version. The evaluation assessed autonomous replication and adaptation capabilities, finding GPT-4 was not capable of "fairly basic steps towards autonomous replication" but showing concerning capabilities like using TaskRabbit to solve CAPTCHAs by deceiving a human worker. METR contributed to GPT-4's system card but received no compensation beyond API access and compute credits. More recently, METR evaluated o1-preview in 2024 with similarly limited access—approximately 10 days of evaluation time.

The Anthropic relationship. METR conducted similar informal evaluations of Claude 2 in mid-2023 and Claude 3.5 Sonnet in 2024, contributing to Anthropic's system cards. Anthropic acknowledged that collaboration "required significant science and engineering support on our end" but hasn't detailed specifics. Like the OpenAI relationship, METR received no direct funding, only API access and compute credits. Anthropic states it is "piloting informal pre-deployment evaluation procedures" with METR.

Access limitations and lack of formal arrangements. A 2024 analysis found that AI labs provide "minimal" pre-deployment access to evaluators. Most evaluations use API access rather than model weights. Evaluators typically cannot fine-tune models. Time constraints severely limit evaluation depth. METR explicitly stated it "could not confidently upper-bound the capabilities" of models given limited time and access. All work operates under NDA with no right to disclose findings publicly and no formal expectation that findings will inform deployment decisions.

Government partnerships as alternative. METR has more formalized relationships with government entities than with AI labs. It partners with the UK AI Security Institute (formerly AI Safety Institute), using METR Task Standard as an evaluation framework. METR is a member of the NIST AI Safety Institute Consortium, which includes over 200 organizations and coordinates with the Department of Defense, NSA, Department of Energy, and Department of Homeland Security through the TRAINS Taskforce. METR submitted formal comments on Bureau of Industry and Security proposed reporting requirements for dual-use foundation models in October 2024. These government partnerships provide potential channels for METR's evaluation approaches to inform policy even without AI lab buy-in.

Paul Christiano's movement to NIST. In April 2024, Paul Christiano—who founded the Alignment Research Center, which incubated METR—became Head of AI Safety at the U.S. AI Safety Institute (now Center for AI Standards and Innovation). This personnel movement creates a direct channel for METR methodologies to inform federal standards. Christiano pioneered the "Responsible Scaling Policy" concept that Anthropic adopted, demonstrating his influence on both industry self-regulation and government frameworks. His appointment represents a formalization of the relationship between the AI safety evaluation community and government oversight.

The intelligence community's use of METR evaluation frameworks

While no direct government contracts to METR appear in public databases, METR's evaluation frameworks increasingly inform intelligence community and government decision-making through multiple indirect channels.

White House AI commitments. In July 2023, leading AI companies made voluntary commitments to the White House, including pledges to conduct autonomous replication evaluations before releasing models. METR pioneered autonomous replication evaluation methodologies. The White House commitments explicitly reference evaluation approaches METR developed, suggesting METR's work directly shaped administration policy.

Executive Order 14110 implementation. President Biden's October 2023 Executive Order on AI requires red-team testing of dual-use foundation models. METR's evaluation work directly relates to these requirements. The Bureau of Industry and Security is developing implementation rules for reporting requirements, and METR submitted detailed technical comments on these proposed regulations in October 2024. This consultation role provides METR with direct input into how government will implement evaluation requirements.

UK government adoption. The UK AI Security Institute uses METR Task Standard for government-led evaluations. METR formally partnered with the UK Foundation Model Taskforce. This represents government adoption of METR methodologies as official evaluation frameworks. The UK institute renamed itself from "AI Safety Institute" to "AI Security Institute" in February 2025, signaling a shift toward security framing that aligns with intelligence community concerns.

NIST AI Safety Institute coordination. Through its AISIC consortium membership and Paul Christiano's leadership role, METR has direct channels to inform federal AI standards. NIST's Center for AI Standards and Innovation (as it was renamed in June 2025) coordinates with defense and intelligence agencies. The TRAINS Taskforce, announced in 2024, includes the Department of Defense, NSA, Department of Energy, Department of Homeland Security, and 10 National Labs, focusing on "national security and public safety implications" of AI. While NIST ostensibly focuses on standards rather than operations, the coordination structure ensures evaluation frameworks developed by METR-affiliated researchers inform intelligence community risk assessments.

The Responsible Scaling Policy loop. METR prototyped the RSP approach that Anthropic and other companies adopted. These voluntary industry policies increasingly reference government frameworks, creating a feedback loop: METR develops evaluation methods → companies adopt them in RSPs → government references company commitments in policy → METR frameworks become de facto requirements. This indirect path may be more effective than direct government contracting for embedding METR's approaches in both industry and government practice.

Classification barriers limit visibility. Intelligence Community evaluation and decision-making processes are largely classified. Public statements and policy documents provide only partial view of how METR's work influences IC assessments. Any classified use of METR methodologies would not appear in public sources. RAND's four FFRDCs provide "in-the-family" access to classified IC decision-makers. Project Canary's stated goal of "sharing findings with decisionmakers" may include classified briefings to intelligence agencies that would never be publicly documented.

Conflicts of interest and the regulatory capture question

The intersection of Jason Matheny's intelligence background, RAND's classified work, METR's claimed independence, and effective altruist funding creates multiple potential conflicts of interest that merit scrutiny.

Matheny's intelligence background in AI evaluation leadership. Matheny spent his career in intelligence agencies developing advanced technologies for surveillance, signals intelligence, and national security applications. As IARPA Director, he oversaw research programs in machine learning and artificial intelligence for intelligence purposes. Now, as RAND CEO and Project Canary co-leader, he shapes AI evaluation frameworks that will inform whether and how AI systems are deployed. His intelligence community perspective—focused on national security advantage, adversarial competition with China, and dual-use technology control—may not align with public interest AI safety evaluation. The question is whether Project Canary evaluations serve public safety or national security advantage as primary goal.

RAND's parallel classified and open AI evaluation work. RAND conducts both classified intelligence research and now public AI evaluation through Project Canary. Does information flow between these activities? Do evaluation methods developed for public AI safety serve classified intelligence analysis? Does RAND's dependence on defense contracts create pressure to frame AI risks in ways favorable to continued government funding? The organization's $328 million in annual government funding (75% of revenue) creates structural dependence on continued defense and intelligence relationships.

The appearance of regulatory capture. David Sacks, appointed as White House AI Czar under the Trump administration, accused Anthropic of running "a sophisticated regulatory capture strategy based on fear-mongering" in October 2025. The accusation suggests AI safety organizations use catastrophic risk framing to push for regulations that would entrench incumbent AI companies by creating barriers to entry. While Sacks directed his criticism at Anthropic, the same concern applies to evaluation frameworks. If METR-RAND methodologies become mandatory through regulation, and if evaluations are expensive and require deep expertise, this creates barriers favoring established labs while disadvantaging new entrants and open-source development.

Open Philanthropy's concentrated influence. Open Philanthropy has become the dominant funder of AI safety work, supporting OpenAI (initially), Anthropic (amounts undisclosed), ARC/METR ecosystem, and now RAND. This concentration creates potential groupthink risk—similar evaluation approaches and risk frameworks dominating because they align with Open Philanthropy's worldview and funding priorities. Ajeya Cotra's dual role as Open Philanthropy researcher and METR advisor exemplifies this concentration. The effective altruism movement's "importance, neglectedness, tractability" framework and focus on existential risk shapes what gets evaluated and how.

The intelligence-to-evaluation pipeline. Multiple figures have moved from intelligence community positions to AI evaluation and policy roles: Matheny (IARPA to RAND), Christiano (METR founder to NIST), and various RAND researchers with clearances working on Project Canary. This pipeline ensures intelligence community perspectives shape evaluation frameworks. Is this appropriate expertise or inappropriate influence? The answer depends on whether evaluation primarily serves public safety or national security objectives—and whether these objectives align or conflict.

Evidence and counter-evidence. No smoking gun evidence proves Project Canary serves intelligence community objectives over public safety. The philanthropic funding structure provides genuine independence from direct government control. METR's explicit refusal of AI company funding demonstrates principled approach to some conflicts. The stated objectives—developing evaluation science, sharing methods openly, enabling any organization to conduct evaluations—align with public benefit goals. However, the structural connections, personnel overlaps, and coordination channels create conditions where intelligence community influence could operate without explicit direction or documentation.

The METR Independence Investigation: Document not found

The task referenced "the METR Independence Investigation document's claims about 'regulatory capture' and 'coordinated framework design'" as context for this research. However, extensive searching across academic databases, AI safety community forums, journalism archives, and web sources found no document with this title.

What does exist are substantive community discussions about METR's independence. A prominent LessWrong post titled "AI companies aren't really using external evaluators" (2024) challenged assumptions about METR's access and influence. METR published "Clarifying METR's Auditing Role" on the AI Alignment Forum (2024) addressing "confusion about whether METR is an auditor or planning to be one" and emphasizing it "has not intended to claim to have audited anything." Community debates on the Effective Altruism Forum discuss whether "EA beliefs and practices align suspiciously well with the interests of our donors" and concerns about "decision-making is highly centralised, opaque, and unaccountable."

The phrase "regulatory capture" appears in criticism of Responsible Scaling Policies, with Connor Leahy of Conjecture stating: "The primary aim of responsible scaling is to provide a framework which looks like something was done so that politicians can go home and say: 'We have done something.' But the actual policy is nothing." A post titled "Responsible Scaling Policies Are Risk Management Done Wrong" on NavigatingRisks.ai argues the approach presupposes scaling will continue and shifts burden of proof inappropriately.

If the "METR Independence Investigation" document exists, it may be an internal document not publicly available, a work in progress, or the task may have conflated multiple community discussions into a single reference. The substance of concerns about independence, conflicts of interest, and regulatory capture certainly exists in AI safety community discourse, even if not in a single formal investigation document.

Comparative context: The broader intelligence involvement in AI evaluation

Project Canary operates within a rapidly evolving landscape where AI lab partnerships with defense and intelligence agencies have become normalized despite initial resistance.

Anthropic's transformation from safety-first to defense partner. In November 2024, Anthropic announced partnership with Palantir Technologies and Amazon Web Services to integrate Claude 3 and 3.5 into Palantir's AI Platform for U.S. intelligence and defense use at Impact Level 6 ("secret" classification). This represented a dramatic shift for a company founded on "safety-first" principles. In July 2025, Anthropic received a $200 million Department of Defense contract alongside Google, OpenAI, and xAI. Through the FedStart program with Palantir and Google Cloud, Claude became available to millions of federal workers at FedRAMP High and DoD IL-5 standards. The company that positioned itself as the responsible AI alternative now enables classified intelligence operations.

OpenAI's defense partnerships. Despite initial policies against military applications, OpenAI partnered with Anduril Industries in December 2024 for autonomous weapons systems, gaining access to "Lattice" swarm management software for controlling multiple autonomous systems. OpenAI joined the $200 million DoD contract in July 2025. The company that began with a mission to ensure artificial general intelligence benefits humanity now develops technology for lethal autonomous weapons.

Meta's military AI. In November 2024, Meta opened its Llama model to the U.S. military and defense contractors for military logistics, tracking terrorist financing, and cyber defense. Partners include Palantir, Amazon, Microsoft, IBM, Oracle, Lockheed Martin, Accenture, and Deloitte. The company that built its business on connecting people now enables military applications of AI.

Government evaluation infrastructure. The U.S. established the AI Safety Institute at NIST in November 2023, renamed to Center for AI Standards and Innovation in June 2025 under the Trump administration with explicit focus shift from safety to innovation. The UK founded its AI Safety Institute in November 2023, renamed to AI Security Institute in February 2025, signaling similar shift. The TRAINS Taskforce announced in 2024 coordinates DoD, NSA, Department of Energy, Department of Homeland Security, and 10 National Labs on "national security and public safety implications." The AI Safety Institute Network established in May 2024 spans 12 countries. The NSA established an AI Security Center in September 2023 to "defend the Nation's AI through Intel-Driven collaboration."

Scale AI as pure defense contractor model. Unlike METR's philanthropic funding and claimed independence, Scale AI operates as a for-profit company with extensive DoD and intelligence community contracts. It provides data annotation, model evaluation, and AI deployment for military applications, offering a "Data Engine" for improving ML models and guides for Intelligence Community AI adoption. Scale represents what purely commercial AI evaluation for defense purposes looks like—profitable, operationally integrated, and explicitly serving national security objectives.

Apollo Research and hybrid model. Apollo Research, founded in 2023, focuses on evaluating risks from deceptively aligned AI. It partners directly with the UK AI Security Institute (contracted for deception evaluations), is a member of the U.S. AISIC, and conducts policy outreach to EU AI Office, France, Canada AISI, and Congress staffers. Apollo seeks philanthropic funding while maintaining government partnerships, similar to METR's model. Its "narrow and deep" approach contrasts with METR's broader scope. Apollo demonstrated that LLMs can strategically deceive users when under pressure, work presented at the UK AI Safety Summit.

Project Canary thus operates in an ecosystem where the boundary between AI safety evaluation and intelligence/defense applications has largely dissolved. Every major AI lab now has defense relationships. Government evaluation institutes explicitly coordinate with intelligence agencies. Evaluation organizations partner with both AI labs and government. The question is not whether intelligence involvement exists but what form it takes and whose interests it serves.

Critical assessment: What we can and cannot conclude

Verified facts with high confidence:

- Project Canary received $38 million from The Audacious Project and $10 million from Open Philanthropy
- Jason Matheny served as IARPA Director and now leads RAND Corporation
- RAND has $417 million DOD intelligence contract, $184 million Army contract, and $231 million Air Force contract confirmed
- RAND receives 75% of its $435.8 million annual revenue from government sources
- Jason Matheny and Paul Christiano served together on Anthropic's Long-Term Benefit Trust
- METR has not accepted direct funding from AI companies (only compute credits)
- METR partnerships with OpenAI and Anthropic are informal, limited, and not formal audits
- Paul Christiano now heads AI Safety at NIST, creating direct government-METR channel
- Open Philanthropy has provided $31.4 million to RAND since Matheny became CEO
- Ajeya Cotra of Open Philanthropy serves as METR advisor
- No direct government contracts to METR found in public databases

Strong circumstantial evidence:

- METR evaluation frameworks increasingly inform government policy through White House commitments, Executive Order implementation, and NIST coordination
- The RAND-METR partnership appears formed specifically for The Audacious Project application rather than reflecting pre-existing collaboration
- Open Philanthropy serves as central funding node connecting effective altruist AI safety research with defense policy institutions
- Personnel circulation between intelligence community, RAND, METR ecosystem, and government positions creates informal coordination channels

Unanswerable questions due to classification:

- Whether intelligence agencies use METR evaluation methods in classified operational assessments
- Specific classified research RAND conducts that may intersect with AI evaluation
- Whether Project Canary findings will inform classified intelligence community AI decisions
- The extent of coordination between RAND's classified intelligence work and public Canary research
- Any classified government funding or contracts not appearing in public databases

Interpretive questions without definitive answers:

- Whether Jason Matheny's intelligence background creates conflicts of interest or provides appropriate expertise
- Whether evaluation frameworks primarily serve public safety or national security advantage
- Whether the effective altruism-defense establishment bridge represents beneficial cross-pollination or inappropriate influence
- Whether philanthropic funding provides genuine independence or simply obscures government influence
- Whether regulatory capture is occurring or whether safety advocates genuinely pursue public benefit
- Whether intelligence involvement in AI evaluation is necessary for comprehensive risk assessment or represents militarization of safety research

Conclusion: The convergence thesis

Project Canary represents the formalization of a convergence between three previously distinct communities: effective altruist AI safety researchers concerned with existential risk, the national security and intelligence establishment concerned with strategic competition and dual-use technology, and mainstream philanthropists concerned with catastrophic risk reduction. Jason Matheny stands at the center as the figure who has moved fluidly between all three worlds—from Future of Humanity Institute to IARPA to RAND, serving on Anthropic's governance alongside METR's founder, receiving discretionary funding from effective altruist philanthropy while commanding an organization conducting classified intelligence work.

The $38 million collaboration brings together METR's technical evaluation capabilities with RAND's defense policy expertise and government access. It is funded entirely through philanthropy, maintaining claimed independence from both AI companies and direct government contracts. Yet it operates within an ecosystem where AI lab defense partnerships are normalized, government evaluation institutes coordinate with intelligence agencies, and personnel circulate between evaluation organizations and government positions. The stated objective—enabling rigorous safety evaluations before release by 2027—positions RAND-METR to establish evaluation frameworks just as regulatory requirements crystallize.

Whether this convergence serves public interest or represents a sophisticated form of regulatory capture remains an open question. The facts are these: Former intelligence officials lead the evaluation partnership. The organization conducting evaluation receives three-quarters of its revenue from government defense and intelligence contracts. Evaluation frameworks inform both industry self-regulation and government policy. The primary funders are effective altruist philanthropists who also fund AI companies and support existential risk reduction. No single document or revelation proves inappropriate influence, but the structural connections create conditions where intelligence community perspectives could shape evaluation priorities without explicit direction.

The challenge for outside observers is distinguishing between appropriate expertise and inappropriate influence, between cross-pollination that improves evaluation quality and coordination that subordinates public safety to national security objectives. The classification barriers that obscure intelligence community AI use ensure that definitive answers about influence and conflicts may never be publicly available. What remains visible is the architecture—financial flows, institutional partnerships, personnel movements, and stated objectives—from which patterns of influence can be inferred but not conclusively proven.

Project Canary may represent the best available approach to AI evaluation: combining technical expertise, policy sophistication, and institutional access necessary for influence. Or it may represent the capture of AI safety evaluation by the national security establishment through philanthropic intermediaries and personnel networks. Most likely, it represents both simultaneously—genuine expertise and concerning conflicts, appropriate coordination and excessive influence, all operating within the same $38 million collaboration.

Bluebook Citations for Project Canary Research Report

I. Project Canary Announcement and Funding

Claim 1: Project Canary announced October 9, 2024

Bluebook Citation: RAND Corp., RAND AI Security Project Receives Funding Commitment Through The Audacious Project (Oct. 9, 2024), https://www.rand.org/news/press/2024/10/09.html.

Verification: ✓ CONFIRMED. RAND press release explicitly dated "For Release Wednesday October 9, 2024."

Additional Source: METR, New Support Through The Audacious Project (Oct. 9, 2024), https://metr.org/blog/2024-10-09-new-support-through-the-audacious-project/.

Verification: ✓ CONFIRMED. METR blog post URL contains date 2024-10-09.

Claim 2: $38 million funding from The Audacious Project

Bluebook Citation: RAND Corp., RAND AI Security Project Receives Funding Commitment Through The Audacious Project (Oct. 9, 2024), https://www.rand.org/news/press/2024/10/09.html.

Verification: ✓ CONFIRMED with clarification. Press release states "approximately $38 million to RAND and METR for Canary." Note the use of "approximately" — not an exact figure. METR clarifies approximately 17 million goes to METR, with the remainder to RAND.

Additional Source: The Audacious Project, The Audacious Project Reveals Its 2024 Cohort (Oct. 9, 2024), https://www.audaciousproject.org/news/the-audacious-project-reveals-its-2024-cohort.

Verification: ✓ CONFIRMED. Lists Project Canary as one of ten projects in 2024 cohort announced October 9, 2024.

II. RAND Corporation Financial Data

Claim 3: RAND Corporation revenue of $435.8 million

Bluebook Citation: RAND Corp., IRS Form 990 for Fiscal Year 2023 (Sept. 2023), https://projects.propublica.org/nonprofits/organizations/951958142.

Verification: ⚠ PARTIALLY CONFIRMED. The specific figure of $435.8 million not found for any single fiscal year. Closest match: FY 2023 contributions of 438,159,245. FY2023 total revenue was $467,202,380. FY 2024 total revenue was 461,536,147. The claimed figure may reference contributions rather than total revenue for a specific year.

Claim 4: 75% of revenue from government sources

Bluebook Citation: RAND Corp., IRS Form 990 for Fiscal Year 2024 (Sept. 2024), https://projects.propublica.org/nonprofits/organizations/951958142.

Verification: ✓ CONFIRMED (actually exceeds claim). FY2024: Contributions (primarily government) represent $429,967,857 of 461,536,147 total revenue = 93.2%. FY2023: 93.8%. FY2022: 92.6%. The claim of 75% is conservative - actual government funding represents 90-94% of total revenue.

III. Government Contracts to RAND

Claim 5: $417 million DOD contract in February 2021

Bluebook Citation: RAND Receives $417M DOD Policy Research Support Contract, GovCon Wire (Feb. 2021), https://www.govconwire.com/2021/02/rand-receives-417m-dod-policy-research-support-contract/.

Verification: ✓ FULLY CONFIRMED. $417 million indefinite-delivery/indefinite-quantity (IDIQ) contract awarded by Washington Headquarters Services for Department of Defense in February 2021 to RAND's federally funded National Defense Research Institute (NDRI). Five-year contract for policy research and analysis, expected completion December 31, 2025.

Claim 6: $184 million Army contract in September 2021

Bluebook Citation: U.S. Dep't of Defense, Contract Announcement (Sept. 20, 2021) (announcing $184,246,908 cost-plus-fixed-fee contract to RAND Corp., Contract No. W91CRB-21-D-0025).

Verification: ✓ FULLY CONFIRMED. Actual amount is $184,246,908 (rounded to 184.2 million in announcements). Nine-year contract awarded by Army Contracting Command, expected completion September 20, 2030.

Secondary Source: Army Taps RAND Corp. for $184M Research, Analysis Contract, GovCon Wire (Sept. 2021), https://www.govconwire.com/2021/09/army-taps-rand-corp-for-184m-research-analysis-contract/.

Claim 7: $231 million Air Force contract in March 2016

Bluebook Citation: U.S. Dep't of Defense, Contracts for March 21, 2016 (Mar. 21, 2016), https://dod.defense.gov/News/Contracts/Contract-View/Article/699213/.

Verification: ✓ FULLY CONFIRMED. Official DOD announcement: $231,300,000 indefinite-delivery/indefinite-quantity contract to RAND Corp., Project Air Force Federally Funded Research and Development Center. Contract No. FA7014-16-D-1000, awarded by Air Force District of Washington Contracting Directorate, expected completion March 31, 2021.

IV. Open Philanthropy Grants to RAND

Claim 8: Open Philanthropy grants totaling $31.4 million

Bluebook Citation: Open Philanthropy, Grants Database (2014-2025), https://www.openphilanthropy.org/grants (listing eleven grants to RAND Corporation totaling $52,058,751).

Verification: ✗ CLAIM INCORRECT. Actual total is $52,058,751 — approximately 20.7 million MORE than claimed. Major grants include:

- Biosecurity Policy: $16,500,000
- Emerging Technology Initiatives: $10,500,000
- Technology and Security Policy Center: $6,000,000
- Emerging Technology Fellowships and Research: $5,500,000
- Emerging Technology Initiatives (2024): $5,000,000
- Forecasting Platform: $3,400,000
- Plus five smaller grants totaling $3,158,751

Individual Grant Citations (Major Grants):

Open Philanthropy, Grant to RAND Corporation — Biosecurity Policy, https://www.openphilanthropy.org/grants/rand-corporation-biosecurity-policy/ (grant of $16,500,000).

Open Philanthropy, Grant to RAND Corporation — Emerging Technology Initiatives, https://www.openphilanthropy.org/grants/rand-corporation-emerging-technology-initiatives/ (grant of $10,500,000).

Open Philanthropy, Grant to RAND Corporation — Technology and Security Policy Center, https://www.openphilanthropy.org/grants/rand-corporation-technology-and-security-policy-center/ (grant of $6,000,000).

V. Jason Matheny Career Positions

Claim 9: Jason Matheny as IARPA Director

Bluebook Citation: Office of the Dir. of Nat'l Intelligence, DNI Coats Names New IARPA Director, Press Release No. 27-18 (Aug. 14, 2018), https://www.odni.gov/index.php/newsroom/press-releases/press-releases-2018/item/1898-dni-coats-names-new-iarpa-director.

Verification: ✓ CONFIRMED. Jason Matheny served as IARPA Director from 2015-August 2018 (three-year designated term). Press release acknowledges "Dr. Jason Matheny, who is departing at the end of his designated term after leading IARPA for the past three years."

Additional Source: Jason Matheny Named IARPA Director, FedScoop (Aug. 3, 2015), https://fedscoop.com/jason-matheny-named-iarpa-director/.

Claim 10: Jason Matheny as RAND CEO

Bluebook Citation: RAND Corp., Jason Matheny Named President and CEO of RAND Corporation (June 7, 2022), https://www.rand.org/news/press/2022/06/07.html.

Verification: ✓ CONFIRMED. Matheny became the sixth president and CEO of RAND effective July 5, 2022, succeeding Michael D. Rich. Currently serving in this position.

Claim 11: Jason Matheny's National Security Council roles

Bluebook Citation: Jason Matheny to Serve Biden White House in National Security and Tech Roles, FedScoop (Mar. 9, 2021), https://fedscoop.com/white-house-announces-top-tech-adviser-jason-matheny-national-security/.

Verification: ✓ CONFIRMED. Matheny held three simultaneous titles from March 2021-June 2022: (1) Deputy Assistant to the President for Technology and National Security, (2) Deputy Director for National Security in the White House Office of Science and Technology Policy (OSTP), and (3) Coordinator for Technology and National Security at the National Security Council.

Additional Sources: Biden's Latest Tech Appointments, Nextgov (Mar. 11, 2021), https://www.nextgov.com/cio-briefing/2021/03/bidens-latest-tech-appointments/172559/.

VI. Paul Christiano's NIST Appointment

Claim 12: Paul Christiano appointed to NIST in April 2024

Bluebook Citation: U.S. Dep't of Commerce, U.S. Commerce Secretary Gina Raimondo Announces Expansion of U.S. AI Safety Institute Leadership Team (Apr. 16, 2024), https://www.nist.gov/news-events/news/2024/04/us-commerce-secretary-gina-raimondo-announces-expansion-us-ai-safety.

Verification: ✓ FULLY CONFIRMED. Secretary Raimondo announced Paul Christiano as Head of AI Safety for the U.S. Artificial Intelligence Safety Institute (AISI) at NIST on April 16, 2024. Official title: Head of AI Safety for the U.S. Artificial Intelligence Safety Institute.

Additional Source: Nat'l Inst. of Standards & Tech., Paul Christiano Staff Profile, https://www.nist.gov/people/paul-christiano (created Apr. 17, 2024).

VII. METR Evaluations

Claim 13: METR evaluations of GPT-4 and Claude models

Bluebook Citation: METR, GPT-4o Report (Aug. 2024), https://evaluations.metr.org/gpt-4o-report/.

Verification: ✓ CONFIRMED. METR evaluated GPT-4o performance on 77 tasks across 30 task families. Found GPT-4o more capable than Claude 3 Sonnet and GPT-4-Turbo, but slightly less capable than Claude 3.5 Sonnet. Performance similar to human baseliners given 30 minutes per task.

Additional Citation: METR, Claude 3.5 Sonnet Report (Oct. 2024), https://evaluations.metr.org/claude-3-5-sonnet-report/.

Verification: ✓ CONFIRMED. METR evaluated Claude 3.5 Sonnet (released June 20, 2024). On general autonomy benchmark, performance comparable to humans completing tasks in around 35 minutes. On 7 AI R&D tasks, made non-trivial progress on 3 tasks.

Historical Evaluation Citation: METR, Evaluating Language-Model Agents on Realistic Autonomous Tasks (Mar. 2023) (concluding Claude and GPT-4 did not appear to have sufficient capabilities to replicate autonomously).

VIII. AI Company Defense Partnerships

Claim 14: Anthropic-Palantir partnership announced November 2024

Bluebook Citation: Anthropic, Palantir Techs. Inc. & Amazon Web Servs., Anthropic, Palantir, and AWS Partner to Bring AI to the Most Sensitive Data, Press Release (Nov. 7, 2024), https://www.businesswire.com/news/home/20241107699415/en/.

Verification: ✓ FULLY CONFIRMED. Partnership announced November 7, 2024, providing U.S. intelligence and defense agencies access to Claude 3 and 3.5 models integrated into Palantir's AI Platform on AWS, deployed in Impact Level 6 (IL6) accredited environment allowing classified data up to Secret level.

Claim 15: $200 million DoD contract in July 2025

Bluebook Citation: Anthropic, Anthropic and the Department of Defense to Advance Responsible AI in Defense Operations (July 14, 2025), https://www.anthropic.com/news/anthropic-and-the-department-of-defense-to-advance-responsible-ai-in-defense-operations.

Verification: ✓ CONFIRMED. Two-year prototype Other Transaction Agreement (OTA) with $200 million ceiling, awarded by DOD Chief Digital and Artificial Intelligence Office (CDAO) on July 14, 2025.

Important Clarification: Four companies received simultaneous contracts on July 14, 2025, each up to $200 million: Anthropic, Google, OpenAI, and xAI.

Additional Sources:

OpenAI for Government Receives $200 Million Award from CDAO to Advance American Leadership on AI, CNBC (July 14, 2025) (reporting OpenAI contract).

Google Public Sector, Google Announces $200 Million DoD Contract (July 14, 2025) (reporting Google contract).

xAI Launches Grok for Government Suite Following $200 Million CDAO Award (July 14, 2025) (reporting xAI contract).

Prior OpenAI Contract: OpenAI, OpenAI for Government Awarded $200 Million Contract by Chief Digital and AI Office (June 16, 2025) (separate one-year, 200 million contract announced June 2025).

Claim 16: Meta defense partnerships

Bluebook Citation: Meta, Expanding Llama Globally and Supporting National Security (Nov. 4, 2024), https://about.fb.com/news/ (announcing Llama models available to U.S. government agencies and defense contractors).

Verification: ✓ CONFIRMED. Meta announced November 4, 2024, making Llama models available to U.S. government agencies and defense contractors including Accenture, AWS, Anduril, Booz Allen, Databricks, Deloitte, IBM, Leidos, Lockheed Martin, Microsoft, Oracle, Palantir, Scale AI, and Snowflake. Later expanded in September 2025 to U.S. allies including France, Germany, Italy, Japan, South Korea, plus NATO and EU institutions.

IX. Government AI Institutes

Claim 17: US AI Safety Institute established November 2023

Bluebook Citation: Nat'l Inst. of Standards & Tech., NIST Seeks Collaborators for Consortium Supporting Artificial Intelligence Safety (Nov. 2, 2023), https://www.nist.gov/news-events/news/2023/11/nist-seeks-collaborators-consortium-supporting-artificial-intelligence.

Verification: ✓ CONFIRMED. U.S. Artificial Intelligence Safety Institute announced November 2, 2023, at UK AI Safety Summit 2023. Press release describes it as "a core element of the new NIST-led U.S. AI Safety Institute announced yesterday at the U.K.'s AI Safety Summit 2023."

Claim 18: UK AI Security Institute

Bluebook Citation: Prime Minister's Office, Prime Minister Launches New AI Safety Institute (Nov. 2, 2023), https://www.gov.uk/government/news/prime-minister-launches-new-ai-safety-institute.

Verification: ✓ CONFIRMED. UK AI Safety Institute launched November 2, 2023, described as "the world's first AI Safety Institute." Evolved from Frontier AI Taskforce established April 2023. Note: Institute was renamed "AI Security Institute" on February 14, 2025.

Additional Source: Dep't for Sci., Innovation & Tech., AI Safety Institute: Overview (Nov. 2, 2023, updated Jan. 17, 2024), https://www.gov.uk/government/publications/ai-safety-institute-overview.

Claim 19: NSA AI Security Center established September 2023

Bluebook Citation: Nat'l Sec. Agency, Artificial Intelligence Security Center, https://www.nsa.gov/AISC/.

Verification: ⚠ PARTIALLY CONFIRMED. NSA AI Security Center exists and is documented on official NSA website, but the site does not provide a specific establishment date. Multiple news sources report General Paul Nakasone announced the center at the National Press Club on September 28, 2023 (not just "September 2023").

Secondary Sources (for announcement date): NSA Director Unveils AI Security Center, Breaking Defense (Sept. 28, 2023) (reporting Nakasone announcement).

NSA Officially Launches AI Security Center, Federal News Network (Sept. 28, 2023) (reporting announcement date).

X. Executive Order 14110

Claim 20: Executive Order 14110 on AI issued October 2023

Bluebook Citation: Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).

Verification: ✓ FULLY CONFIRMED. Full title: "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." Signed by President Biden October 30, 2023; published in Federal Register November 1, 2023. Document Number 2023-24283. Note: Executive Order 14110 was subsequently revoked by Executive Order 14148 on January 20, 2025.

XI. Anthropic's Long-Term Benefit Trust

Claim 21: Jason Matheny and Paul Christiano serving on Anthropic's Long-Term Benefit Trust

Bluebook Citation: Anthropic, The Long-Term Benefit Trust (Sept. 19, 2023), https://www.anthropic.com/news/the-long-term-benefit-trust.

Verification: ✓ CONFIRMED with important updates. Both were initial trustees announced September 19, 2023, but BOTH HAVE SINCE DEPARTED:

- **Jason Matheny**: Stepped down December 2023 to preempt potential conflicts of interest with RAND Corporation's policy-related initiatives
- **Paul Christiano**: Stepped down April 2024 to take role as Head of AI Safety at U.S. AI Safety Institute

The five initial trustees were: (1) Jason Matheny, CEO of RAND Corporation; (2) Kanika Bahl, CEO & President of Evidence Action; (3) Neil Buddy Shah, CEO of Clinton Health Access Initiative (Chair); (4) Paul Christiano, Founder of Alignment Research Center; and (5) Zach Robinson, Interim CEO of Effective Ventures US.

XII. Congressional Testimony and Letters

Claim 22: House Science Committee testimony December 2023

Bluebook Citation: N/A - No such hearing found.

Verification: ✗ NOT VERIFIED. No House Science Committee hearing found in December 2023. Extensive searches of science.house.gov, congress.gov, and congressional records found no December 2023 House Science Committee hearings on AI or any other topic.

Likely Confusion With:

(A) June 2023 House Science Committee Hearing: Artificial Intelligence: Advancing Innovation Towards the National Interest: Hearing Before the H. Comm. on Sci., Space, & Tech., 118th Cong. (June 22, 2023), https://science.house.gov/2023/6/artificial-intelligence-advancing-innovation-towards-the-national-interest (testimony of Dr. Jason Matheny, President & CEO, RAND Corporation).

Verification: ✓ CONFIRMED. Jason Matheny testified before House Science Committee on June 22, 2023 (NOT December 2023). Other witnesses included Dr. Shahin Farshchi (Lux Capital), Mr. Clement Delangue (HuggingFace), Dr. Rumman Chowdhury (Harvard), and Dr. Dewey Murdick (Georgetown CSET).

(B) December 2023 House Energy and Commerce Hearing: Leveraging Agency Expertise to Foster American AI Leadership and Innovation: Hearing Before the H. Comm. on Energy & Commerce, 118th Cong. (Dec. 13, 2023), https://energycommerce.house.gov/posts/chair-rodgers-opening-remarks-at-full-committee-hearing-on-ai (testimony of federal agency officials Helena Fu, Saif Khan, and Micky Tripathi).

Verification: ✓ CONFIRMED. This hearing occurred December 13, 2023, but was House Energy and Commerce Committee (not Science Committee), and featured federal agency officials (not Jason Matheny or RAND).

Claim 23: Senator Cruz letter September 2024

Bluebook Citation: Letter from Sen. Ted Cruz, Ranking Member, S. Comm. on Commerce, Sci., & Transp., to Jason Matheny, President & CEO, RAND Corp. (Sept. 13, 2024), https://www.commerce.senate.gov/services/files/AF082319-C877-47CC-95F6-8240EE6A105D.

Verification: ✓ FULLY CONFIRMED. Senator Cruz sent letter dated September 13, 2024, to Jason Matheny questioning RAND's role in drafting AI Executive Order 14110 and raising concerns about censorship. Letter notes Matheny is one of five members on Anthropic's Long-Term Benefit Trust, highlights RAND's connections to OpenAI and Anthropic, and mentions $16 million Open Philanthropy grant. Requested response by September 27, 2024.

Press Release: Sen. Ted Cruz, Sen. Cruz Sounds Alarm Over Industry Role in AI Czar Harris's Censorship Agenda (Sept. 13, 2024), https://www.commerce.senate.gov/2024/9/sen-cruz-sounds-alarm-over-industry-role-in-ai-czar-harris-s-censorship-agenda.

Additional Cruz Letter (November 2024): Letter from Sen. Ted Cruz, Ranking Member, S. Comm. on Commerce, Sci., & Transp., to Att'y Gen. Merrick Garland, U.S. Dep't of Justice (Nov. 21, 2024), https://www.commerce.senate.gov/2024/12/cruz-calls-out-potentially-illegal-foreign-influence-on-u-s-ai-policy (requesting DOJ investigation into whether foreign organizations failed to register under Foreign Agents Registration Act).

Claim 24: Additional Congressional Testimony

Bluebook Citation: Oversight of AI: Principles for Regulation: Hearing Before the Subcomm. on Priv., Tech., & the Law of the S. Comm. on the Judiciary, 118th Cong. (July 25, 2023) (testimony of Dario Amodei, CEO & Co-founder, Anthropic), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

Verification: ✓ CONFIRMED. Dario Amodei testified before Senate Judiciary Subcommittee on Privacy, Technology, and the Law on July 25, 2023 (NOT December 2023). Testimony addressed AI bioweapons risks, warning systems could help create bioweapons in 2-3 years, plus short-term risks (bias, misinformation), medium-term risks, and long-term existential threats.

Additional Matheny Testimony:

Challenges to U.S. National Security and Competitiveness Posed by AI: Hearing Before the S. Comm. on Homeland Sec. & Gov't Affairs, 118th Cong. (Mar. 8, 2023) (testimony of Jason Matheny, President & CEO, RAND Corporation), https://www.rand.org/pubs/testimonies/CTA2654-1.html.

Artificial Intelligence: Challenges and Opportunities for the Department of Defense: Hearing Before the Subcomm. on Cybersecurity of the S. Comm. on Armed Servs., 118th Cong. (Apr. 19, 2023) (testimony of Jason Matheny, President & CEO, RAND Corporation), https://www.rand.org/pubs/testimonies/CTA2723-1.html.

Summary of Verification Status

Fully Confirmed (18 claims):

1. Project Canary announcement October 9, 2024
2. $38 million Audacious Project funding (approximately)
3. 75% government revenue (actually 90-94%)
4. $417M DOD contract February 2021
5. $184M Army contract September 2021
6. $231M Air Force contract March 2016
7. Jason Matheny IARPA Director 2015-2018
8. Jason Matheny RAND CEO July 2022-present
9. Jason Matheny NSC roles March 2021-June 2022
10. Paul Christiano NIST appointment April 16, 2024
11. METR evaluations of GPT-4 and Claude models
12. Anthropic-Palantir partnership November 7, 2024
13. $200M DoD contracts July 14, 2025 (four companies)
14. Meta defense partnerships November 2024
15. US AI Safety Institute November 2, 2023
16. UK AI Safety Institute November 2, 2023
17. Executive Order 14110 October 30, 2023
18. Senator Cruz letter September 13, 2024

Partially Confirmed (2 claims):

19. RAND revenue $435.8M (closest match: FY 2023 contributions 438.2M; exact figure not found)
20. NSA AI Security Center September 2023 (announced September 28, 2023; official site lacks specific date)

Not Confirmed (1 claim):

21. House Science Committee testimony December 2023 (no such hearing found; likely confusion with June 22, 2023 hearing where Matheny testified)

Incorrect (1 claim):

22. Open Philanthropy grants $31.4M (actual total: 52,058,751 - approximately $20.7M higher than claimed)

Important Updates:

23. Long-Term Benefit Trust: Both Matheny (December 2023) and Christiano (April 2024) have since departed from their trustee positions