

EXHIBIT B-31 Matheny-Christiano Anthropic Trustee Documentation.md

Legally Relevant Events for Kirchner v. Johnson

Jason Matheny - Government-to-Private-to-Government Pipeline

Intelligence Community Leadership (2015-2018)

2015 - Appointed Director of Intelligence Advanced Research Projects Activity (IARPA), holding title of Assistant Director of National Intelligence *Relevance: Oversight of advanced technology development for U.S. intelligence community spanning quantum computing, forecasting systems, and open-source intelligence across 500+ universities and research institutions*

2015 - Appointed to lead foundational R&D for National Strategic Computing Initiative

2018-2021 - Served on National Security Commission on Artificial Intelligence *Relevance: Established national security credentials in AI policy*

Think Tank Positioning and EA Strategy (2017-2019)

2017 - Delivered speech to effective altruism community: "Effective Altruism in Government" *Relevance: Explicitly articulated strategy of personnel placement to advance AI safety priorities, stating "there's a lot of low-hanging fruit within government positions that we should be picking as effective altruists" and advocating for influence both "inside" government and "outside" through contractors and think tanks*

January 2019 - Founded Georgetown's Center for Security and Emerging Technology (CSET) with \$55 million grant from Open Philanthropy *Relevance: Open Philanthropy grant documents stated that if Matheny were to depart CSET, they would need to "reconsider our support for, and alignment with, the future plans" of the center*

Biden Administration Service (2021-2022)

March 2021 - June 2022 - Served in Biden White House in three simultaneous roles:

- Deputy Assistant to the President for Technology and National Security
- Deputy Director for National Security at OSTP

- NSC Coordinator for Technology and National Security *Relevance: First person to bridge National Security Council and Office of Science and Technology Policy; advised on keeping U.S. ahead of China in semiconductors and AI*

RAND Corporation Leadership (2022-Present)

July 2022 - Appointed President and CEO of RAND Corporation (nonprofit research organization with ~2,000 employees and over \$350 million annual revenue)

2023 - RAND received over \$15 million from Open Philanthropy *Relevance: RAND researchers identified as "key policy influencers" behind White House AI Executive Order*

March, April, June, August 2023 - Testified before House and Senate committees on AI governance challenges

August 2023 - Published Washington Post op-ed (while serving as trustee) arguing for AI oversight focusing on hardware, training, and deployment supply chain *Relevance: Framework directly relevant to Anthropic's operations*

Anthropic Trustee Tenure (September - December 2023)

September 2023 - Appointed to Anthropic Long-Term Benefit Trust as one of five initial trustees *Relevance: Created immediate tension between advising government on AI regulation while simultaneously serving as trustee of major AI company, just 14 months into RAND leadership*

September 19, 2023 - LTBT and Responsible Scaling Policy announced simultaneously *Relevance: Coordinated launch of governance framework during Matheny's first month as trustee*

September 2023 - Amazon announced **1.25 billion initial investment in Anthropic (part of eventual 4 billion commitment)** *Relevance: Occurred during Matheny's trustee tenure*

October 2023 - Google invested **500 million in Anthropic (with commitment to additional 1.5 billion)** *Relevance: Massive investment creating potential tensions with trust's mission-focused mandate*

November 2023 - Claude 2.1 launched with 200,000-token context window and 2x decrease in hallucination rates *Relevance: Significant capability increase during trustee tenure; no documentation of Matheny's involvement in deployment decision*

December 2023 - Resigned from Anthropic LTBT after just **three months**, explicitly citing need "to preempt any potential conflicts of interest that might arise with RAND Corporation's policy-related initiatives" *Relevance: Proactive resignation demonstrates awareness of conflicts; rapid departure raises questions about vetting process*

Post-Trustee Developments

2024-Present - Continued RAND leadership while maintaining advisory board positions for Department of Commerce, Department of Energy, and Nuclear Threat Initiative

Reports indicate - RAND's warnings about AI risks may conflict with Trump administration's AI policy approach *Relevance: Ongoing tension between RAND's policy positions and political environment*

Paul Christiano - Framework Designer to Trustee to Regulator

OpenAI Period (2017-2021)

January 2017 - Joined OpenAI immediately upon completing PhD

June 12, 2017 - First submission of "Deep Reinforcement Learning from Human Preferences" (foundational RLHF paper), co-authored with Dario Amodei and others *Relevance: Established deep technical collaboration with future Anthropic CEO*

September 2019 - January 2021 - Led OpenAI's Language Model Alignment Team

January 2021 - Left OpenAI; stated reasons: "desire to work on more conceptual/theoretical issues in alignment" and positioning outside corporate labs for policy influence *Relevance: Timing coincided with Dario Amodei and other researchers leaving to found Anthropic*

Alignment Research Center Founding (2021-2023)

April 26, 2021 - Publicly announced founding of Alignment Research Center (ARC), 501(c)(3) nonprofit

March 2022 - ARC received \$265,000 from Open Philanthropy

2022 - Beth Barnes joined ARC from OpenAI to form ARC Evals as incubated project *Relevance: Created evaluation organization that would advise Anthropic on RSP framework*

November 2022 - ARC received

1,250,000two — yeargrant from OpenPhilanthropy * *Relevance : Total funding* 1.515 million from organization whose AI funding is led by Ajeya Cotra (Christiano's wife)*

Early 2023 - ARC Evals conducted pre-deployment evaluation of GPT-4 for OpenAI, including TaskRabbit CAPTCHA test

2023 - ARC Evals conducted Claude evaluations for Anthropic

Anthropic Trustee Tenure (September 2023 - April 2024)

September 2023 - Appointed to Anthropic Long-Term Benefit Trust as one of five initial trustees

September 19, 2023 - COORDINATED ANNOUNCEMENTS:

- Anthropic released RSP v1.0 stating it was designed "in the spirit of" the framework developed by Paul Christiano and ARC Evals
- Anthropic explicitly acknowledged: "We thank ARC Evals for substantial advice and collaboration on the development of our commitments"

- Same day: ARC Evals announced spin-out from ARC as independent organization *Relevance: Christiano simultaneously (1) founded organization that developed RSP conceptual framework; (2) received payment from Anthropic for evaluation services; (3) served as trustee with oversight responsibility; (4) led organization evaluating Anthropic's models*

November 2023 - Claude 2.1 released during Christiano's trustee tenure *Relevance: No public documentation of Christiano's involvement in deployment decision despite RSP requiring trustee consultation*

December 4, 2023 - ARC Evals officially rebranded as METR. Announcement stated: "While we had initially planned for Paul Christiano to be a board member and advisor, **he declined involvement in light of his new role as Head of Safety at the US AI Safety Institute**" *Relevance: Indicates government appointment process already underway by December 2023; declined METR board involvement while remaining Anthropic trustee through April 2024*

March 2024 - Preparation for Claude 3 release (released March 2024) *Relevance: Significant capability increase during trustee tenure*

March 2024 - VentureBeat reported NIST staff revolt over Christiano's anticipated appointment; staff concerns about "ties to the effective altruism movement may jeopardize the AI Safety Institute's objectivity and integrity"

April 2024 - Resigned from Anthropic LTBT after six months to take government position *Relevance: Official statement: "Paul Christiano stepped down in April 2024 to take a new role as the Head of AI Safety at the U.S. AI Safety Institute"*

April 16, 2024 - U.S. Secretary of Commerce Gina Raimondo officially announced Christiano's appointment as Head of AI Safety at U.S. AI Safety Institute

Post-Trustee Government Service

May 2024 - LTBT elected Jay Kreps (Confluent CEO) as first board member - **one month after Christiano's departure** *Relevance: Eight months after trust announcement; timing raises questions about delays*

August 29, 2024 - U.S. AI Safety Institute signed agreements with OpenAI and Anthropic for AI safety research, testing, and evaluation *Relevance: Christiano now in position to oversee companies he previously advised and served as trustee*

May 2025 - LTBT elected Reed Hastings (Netflix co-founder) as second board member *Relevance: Trust now controls 2 of 5 board seats*

June 2025 - U.S. AI Safety Institute renamed "Center for AI Standards and Innovation" (CAISI) under Trump Administration *Relevance: Christiano's current role status and authority uncertain*

The Three-Month Overlap Period (September - December 2023)

Simultaneous Service During Critical Period

September 2023 - Both Matheny and Christiano serving as trustees when:

- LTBT publicly announced (September 2023)
- RSP framework released crediting Christiano's ARC Evals (September 19, 2023)
- Amazon investment announced: \$1.25 billion initial (September 2023)

October 2023 - Both serving when:

- Google invested \$500 million (October 2023)
- LTBT governance structure implemented

November 2023 - Both serving when:

- Claude 2.1 released with significant improvements (November 2023)
- OpenAI board fired and reinstated Sam Altman
- CEO Dario Amodei positioned Anthropic as stable alternative, stating OpenAI situation "could never happen at Anthropic" due to LTBT structure

December 2023 - Matheny resigned, leaving Christiano as sole trustee with AI safety expertise *Relevance: Rapid turnover of both trustees with deep AI safety expertise within first quarter of operation*

Long-Term Benefit Trust Structure and Transparency Issues

Governance Framework

October 2023 - LTBT structure announced combining:

- Delaware Public Benefit Corporation status
- Common law purpose trust holding special Class T stock
- Authority to elect board members: initially 1 of 5 directors, phasing to 3 of 5 (majority control) within four years

Five Initial Trustees (September 2023):

1. Jason Matheny (RAND CEO) - *departed December 2023*
2. Paul Christiano (ARC founder) - *departed April 2024*
3. Neil Buddy Shah (Clinton Health Access Initiative CEO, Chair) - *active*
4. Kanika Bahl (Evidence Action CEO) - *active*
5. Zach Robinson (Effective Ventures US Interim CEO) - *active*

Trustee Compensation: "Modest" compensation with **no equity** in Anthropic - designed to make trustees "financially disinterested"

Limited Transparency Documentation

No Public Records Exist Of:

- Full Trust Agreement showing specific powers and limitations
- Minutes or summaries of trustee meetings
- Documentation of information trustees requested or received
- Records of trustee consultation on RSP changes or model deployments
- Details of board appointee vetting and selection process
- Evidence of any trustee interventions, concerns raised, or decisions blocked

Only Public Documentation:

- Current trustee list and board members on Anthropic website
- Occasional update footnotes
- No systematic reporting or benefit reports despite Delaware PBC status allowing such reports

Board Elections and Authority

May 2024 - Jay Kreps appointed as first LTBT-elected board member (8 months after trust announcement, 1 month after Christiano departure)

May 2025 - Reed Hastings appointed as second LTBT-elected board member

Current Status: LTBT controls 2 of 5 board seats; designed to reach 3 of 5 (majority control) by November 2025

Critical Limitation: Board members elected by LTBT still owe fiduciary duties to all stockholders, not exclusively to trust's public benefit mission *Relevance: Creates fundamental tension when safety concerns conflict with financial interests*

Network Connections and Conflicts of Interest

Financial Relationships

Open Philanthropy Funding Network:

- CSET (Matheny founded): \$55 million
- RAND (Matheny CEO): \$15+ million in 2023
- ARC (Christiano founded): \$1.515 million in 2022

Personal Relationships:

- Paul Christiano married to Ajeya Cotra (leads AI funding at Open Philanthropy)
- Holden Karnofsky (Open Philanthropy co-CEO) married to Daniela Amodei (Anthropic President) *Relevance: Direct financial and personal links across ecosystem*

Effective Altruism Movement Integration

Common Philosophical Framework:

- All three departed/replaced trustees (Matheny, Christiano, Robinson) have EA movement connections
- Shared focus on longtermism and AI existential risk
- Dense network where same individuals appear as co-authors, co-founders, trustees, advisors, regulators

Matheny's 2017 EA Strategy: Explicit advocacy for "inserting personnel who share his AI worldview into government and government contractor roles"

Documented Conflicts and Management Attempts

Matheny Conflicts

Active During Trustee Tenure:

- RAND CEO (July 2022 - present)
- Anthropic LTBT trustee (September - December 2023)
- RAND receiving \$15+ million from Open Philanthropy
- RAND researchers influencing White House AI policy
- Congressional testimony on AI governance

Resolution: Resigned December 2023 after 3 months, citing need to "preempt conflicts"

Christiano Conflicts

Active During Trustee Tenure:

- ARC founder and executive director
- ARC Evals paid consultant to Anthropic on RSP development
- ARC Evals conducting evaluations of Anthropic models
- LTBT trustee with oversight responsibility
- Married to Open Philanthropy AI funding lead

Partial Resolution:

- ARC Evals spun out as independent METR (September 2023)
- Declined METR board involvement (December 2023)
- Resigned from LTBT (April 2024) upon government appointment

Unresolved:

- Designed RSP frameworks he now evaluates as government regulator
- METR continues evaluating companies he oversees
- Personal relationships with company leadership remain

Timeline Summary for Legal Analysis

Pattern of Insufficient Conflict Management

1. **Short Tenures:** Matheny (3 months), Christiano (6 months) - 40% trustee turnover within 8 months
2. **Predictable Conflicts:** Both conflicts were foreseeable at appointment:
 - Matheny: RAND CEO role creating government advisory conflicts
 - Christiano: ARC Evals paid relationship with Anthropic
3. **Expertise Loss:** Departures eliminated both trustees with deep AI safety expertise (technical and policy)
4. **Transparency Failure:** No public record of any specific decisions or actions either trustee made
5. **Coordinated Announcements:** September 19, 2023 same-day release of RSP (crediting Christiano's work) and ARC Evals spin-out demonstrates coordination pattern

Framework Designer → Evaluator → Regulator Pipeline

Christiano's Path:

- Designed RLHF methodology (2017) → Industry standard
- Designed RSP framework concept (2023) → Adopted by Anthropic
- Conducted evaluations through ARC Evals (2023) → For companies using his frameworks
- Served as trustee (2023-2024) → Overseeing company implementing his frameworks
- Government regulator (2024-present) → Setting standards for companies using his frameworks

Unresolved Questions

1. Why appoint Matheny despite foreseeable RAND CEO conflicts?
2. Why not address Christiano-ARC Evals conflict before appointment?
3. What decisions did either trustee actually make during tenure?
4. How were conflicts managed during overlapping service period?
5. Did trustees participate in Claude 2.1 deployment decision?
6. What happens when trust reaches majority board control with diminished expertise?

Comprehensive Bluebook Reference List

Anthropic's Long-Term Benefit Trust Governance Structure

A

Aboutamazon.com, Amazon to Invest Another \$4 Billion in Anthropic, AMAZON (Mar. 27, 2024), <https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment>.

Aboutamazon.com, Amazon Will Invest up to \$4B in Anthropic to Advance Generative AI, AMAZON (Sept. 25, 2023), <https://www.aboutamazon.com/news/company-news/amazon-aws-anthropic-ai>.

Allyn, Bobby, Why CEO Sam Altman's Firing at ChatGPT Maker Open AI Was Bound to Happen, NPR (Nov. 24, 2023), <https://www.npr.org/2023/11/24/1215015362/chatgpt-openai-sam-altman-fired-explained>.

Anthropic, Activating AI Safety Level 3 Protections, ANTHROPIC (May 2025), <https://www.anthropic.com/activating-asl3-report>.

Anthropic, Anthropic Raises **13B Series F** at 183B Post-Money Valuation, ANTHROPIC (Sept. 2, 2025), <https://www.anthropic.com/news/anthropic-raises-series-f-at-usd183b-post-money-valuation>.

Anthropic, Anthropic's Responsible Scaling Policy, ANTHROPIC (Sept. 19, 2023), <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>.

Anthropic, Claude 2, ANTHROPIC (July 11, 2023), <https://www.anthropic.com/news/claude-2>.

Anthropic, Claude's Constitution, ANTHROPIC, <https://www.anthropic.com/news/claudes-constitution>.

Anthropic, Collective Constitutional AI: Aligning a Language Model with Public Input, ANTHROPIC (Oct. 17, 2023), <https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input>.

Anthropic, Golden Gate Claude, ANTHROPIC (May 21, 2024), <https://www.anthropic.com/news/golden-gate-claude>.

Anthropic, Introducing Claude, ANTHROPIC (Mar. 14, 2023), <https://www.anthropic.com/news/introducing-claude>.

Anthropic, Introducing Claude 2.1, ANTHROPIC (Nov. 21, 2023), <https://www.anthropic.com/news/claude-2-1>.

Anthropic, Introducing Claude 4, ANTHROPIC (May 22, 2025), <https://www.anthropic.com/news/claude-4>.

Anthropic, Introducing the Next Generation of Claude, ANTHROPIC (Mar. 4, 2024), <https://www.anthropic.com/news/claude-3-family>.

Anthropic, Introducing Claude 3.5 Sonnet, ANTHROPIC (June 20, 2024), <https://www.anthropic.com/news/claude-3-5-sonnet>.

Anthropic, Jay Kreps Appointed to Anthropic's Board of Directors, ANTHROPIC (May 29, 2024), <https://www.anthropic.com/news/jay-kreps-appointed-to-board-of-directors>.

Anthropic, The Long-Term Benefit Trust, ANTHROPIC (Oct. 2023), <https://www.anthropic.com/news/the-long-term-benefit-trust>.

Anthropic, Mapping the Mind of a Large Language Model, ANTHROPIC (May 2024), <https://www.anthropic.com/research/mapping-mind-language-model>.

Anthropic, Reflections on Our Responsible Scaling Policy, ANTHROPIC, <https://www.anthropic.com/news/reflections-on-our-responsible-scaling-policy>.

Anthropic, Releasing Claude Instant 1.2, ANTHROPIC (Aug. 9, 2023), <https://www.anthropic.com/news/releasing-claude-instant-1-2>.

Anthropic, Responsible Scaling Policy, Version 1.0 (Sept. 19, 2023), <https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf>.

Anthropic, System Card: Claude Opus 4 & Claude Sonnet 4 (May 2025), <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>.

B

Bai, Yuntao et al., Constitutional AI: Harmlessness from AI Feedback, ARXIV (Dec. 15, 2022), <https://arxiv.org/abs/2212.08073>.

Barnes, Beth, Evaluations Project @ ARC is Hiring a Researcher and a Webdev/Engineer, LESSWRONG (Oct. 2022), <https://www.lesswrong.com/posts/svhQMdsefdYFDq5YM/evaluations-project-arc-is-hiring-a-researcher-and-webdev-1>.

Beauchamp, Zack, 4 Days from Fired to Re-Hired: A Timeline of Sam Altman's Ouster from OpenAI, ABC NEWS (Nov. 21, 2023), <https://abcnews.go.com/Business/sam-altman-reaches-deal-return-ceo-openai/story?id=105091534>.

C

Carey, Ryan, Anthropic's Responsible Scaling Policy & Long-Term Benefit Trust, AI ALIGNMENT F. (Oct. 25, 2023), <https://www.alignmentforum.org/posts/6tjHf5ykvFqaNCErH/anthropic-s-responsible-scaling-policy-and-long-term-benefit>.

Christiano, Paul, Announcing the Alignment Research Center, AI ALIGNMENT F. (Apr. 26, 2021), <https://www.alignmentforum.org/posts/3ejHFgQihLG4L6WQf/announcing-the-alignment-research-center>.

Christiano, Paul, My Views on "Doom," LESSWRONG (June 2023), <https://www.lesswrong.com/posts/xWMqsvHapP3nwdSW8/my-views-on-doom>.

Christiano, Paul et al., Deep Reinforcement Learning from Human Preferences, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017) (2017), <https://arxiv.org/abs/1706.03741>.

Christiano, Paul, Buck Shlegeris & Dario Amodei, Supervising Strong Learners by Amplifying Weak Experts, ARXIV (Oct. 2018), <https://arxiv.org/abs/1810.08575>.

Coldewey, Devin, Anthropic is the New AI Research Outfit from OpenAI's Dario Amodei, and It Has \$124M to Burn, TECHCRUNCH (May 28, 2021), <https://techcrunch.com/2021/05/28/anthropic-is-the-new-ai-research-outfit-from-openais-dario-amodei-and-it-has-124m-to-burn/>.

Cotra, Ajeya, Iterated Distillation and Amplification, AI ALIGNMENT (Mar. 4, 2018), <https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616>.

D

DEL. CODE ANN. tit. 8, §§ 361-368 (West 2024).

F

Feiner, Lauren, Former OpenAI Board Member Explains Why CEO Sam Altman Got Fired Before He Was Rehired, CNBC (May 29, 2024), <https://www.cnbc.com/2024/05/29/former-openai-board-member-explains-why-ceo-sam-altman-was-fired.html>.

Fortune Magazine, A.I. Has a '10 or 20% Chance' of Conquering Humanity, Former OpenAI Safety Researcher Warns, FORTUNE (May 3, 2023), <https://fortune.com/2023/05/03/openai-ex-safety-researcher-warns-ai-destroy-humanity/>.

G

Georgetown University, Largest U.S. Center on Artificial Intelligence, Policy Comes to Georgetown, GEORGETOWN U. (Feb. 28, 2019), <https://www.georgetown.edu/news/largest-u-s-center-on-artificial-intelligence-policy-comes-to-georgetown/>.

Georgetown University Center for Security and Emerging Technology, Jason Matheny to Serve Biden White House in National Security and Tech Roles, CSET (Mar. 8, 2021), <https://cset.georgetown.edu/article/jason-matheny-to-serve-biden-white-house-in-national-security-and-tech-roles/>.

Goldman, Sharon, AI and Policy Leaders Debate Web of Effective Altruism in AI Security, VENTUREBEAT (Dec. 11, 2023), <https://venturebeat.com/ai/ai-and-policy-leaders-debate-web-of-effective-altruism-in-ai-security-the-ai-beat/>.

Goldman, Sharon, The Widening Web of Effective Altruism in AI Security, VENTUREBEAT (Nov. 13, 2023), <https://venturebeat.com/ai/the-widening-web-of-effective-altruism-in-ai-security-the-ai-beat/>.

Good Ventures, Center for Security and Emerging Technology — General Support, GOOD VENTURES (Jan. 2021), <https://www.goodventures.org/our-portfolio/grants/center-for-security-and-emerging-technology-general-support/>.

H

Hays, Kali, ChatGPT Parent Company OpenAI Fires CEO Sam Altman, CNN (Nov. 17, 2023), <https://www.cnn.com/2023/11/17/tech/sam-altman-departs-open-ai/index.html>.

Heilweil, Rebecca, OpenAI, Anthropic Enter AI Agreements with US AI Safety Institute, FEDSCOOP (Aug. 29, 2024), <https://fedscoop.com/openai-anthropic-enter-ai-agreements-with-us-ai-safety-institute/>.

I

Irving, Geoffrey, Paul Christiano & Dario Amodei, AI Safety via Debate, ARXIV (May 2, 2018), <https://arxiv.org/abs/1805.00899>.

K

Kahn, Jeremy, Anthropic Hired President Daniela Amodei's Husband to Work on the Company's AI Safety Strategy, FORTUNE (Feb. 13, 2025), <https://fortune.com/2025/02/13/anthropic-hired-president-daniela-amodei-husband-ai-safety-responsible-scaling/>.

L

Lerman, Rachel, Why Sam Altman Was Fired, and Why He Is Back at OpenAI, Explained, WASH. POST (Nov. 22, 2023), <https://www.washingtonpost.com/technology/2023/11/20/openai-sam-altman-ceo-oust/>.

Lunden, Ingrid, UK Drops 'Safety' From Its AI Body, Now Called AI Security Institute, Inks MOU with Anthropic, TECHCRUNCH (Feb. 13, 2025), <https://techcrunch.com/2025/02/13/uk-drops-safety-from-its-ai-body-now-called-ai-security-institute-inks-mou-with-anthropic/>.

M

Matheny, Jason, A Simple Way for the United States to Regulate Powerful AI Models, WASH. POST (Aug. 16, 2023), <https://www.washingtonpost.com/opinions/2023/08/16/ai-danger-regulation-united-states/>.

Matheny, Jason, Advancing Trustworthy Artificial Intelligence, Testimony Before the H. Comm. on Sci., Space, & Tech. (June 22, 2023), https://democrats-science.house.gov/download/dr-matheny_testimony.

Matheny, Jason, Advancing Trustworthy Artificial Intelligence: Addendum, Testimony Before the H. Comm. on Sci., Space, & Tech. (Aug. 1, 2023), <https://www.rand.org/pubs/testimonies/CTA2824-2.html>.

Matheny, Jason, Artificial Intelligence: Challenges and Opportunities for the Department of Defense, Testimony Before the S. Armed Servs. Subcomm. on Cybersecurity (Apr. 19, 2023), <https://www.rand.org/pubs/testimonies/CTA2723-1.html>.

Matheny, Jason, Challenges to U.S. National Security and Competitiveness Posed by AI, Testimony Before the S. Comm. on Homeland Sec. & Governmental Affairs (Mar. 8, 2023), <https://www.rand.org/pubs/testimonies/CTA2654-1.html>.

Matheny, Jason, Effective Altruism in Government, EFFECTIVE ALTRUISM (2017), <https://www.effectivealtruism.org/articles/effective-altruism-in-government-jason-matheny>.

Matheny, Jason, Looking Back on the Big Policy Stories of 2023, RAND CORP. (Dec. 2023), <https://www.rand.org/pubs/commentary/2023/12/looking-back-on-the-big-policy-stories-of-2023.html>.

Matthews, Dylan, The \$1 Billion Gamble to Ensure AI Doesn't Destroy Humanity, VOX (Sept. 25, 2023), <https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2>.

METR, ARC Evals is Now METR, METR BLOG (Dec. 4, 2023), <https://metr.org/blog/2023-12-04-metr-announcement/>.

METR, ARC Evals is Spinning Out from ARC, METR BLOG (Sept. 19, 2023), <https://metr.org/blog/2023-09-19-spin-out-announcement/>.

METR, More Information About the Dangerous Capability Evaluations We Did with GPT-4 and Claude, METR BLOG (Mar. 18, 2023), <https://metr.org/blog/2023-03-18-update-on-recent-evals/>.

Morley, John, David J. Berger \u0026 Amy L. Simmerman, Anthropic Long-Term Benefit Trust, HARV. L. SCH. F. ON CORP. GOVERNANCE (Oct. 28, 2023), <https://corpgov.law.harvard.edu/2023/10/28/anthropic-long-term-benefit-trust/>.

N

NIST, U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI, NAT'L INST. OF STANDARDS \u0026 TECH. (Aug. 29, 2024), <https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research>.

O

OpenAI, GPT-4 System Card, OPENAI (Mar. 23, 2023), <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.

OpenAI, GPT-4 Technical Report, ARXIV (Mar. 15, 2023), <https://arxiv.org/abs/2303.08774>.

Open Philanthropy, Alignment Research Center — General Support, OPEN PHILANTHROPY (Mar. 2022), <https://www.openphilanthropy.org/grants/alignment-research-center-general-support/>.

Open Philanthropy, Alignment Research Center — General Support, OPEN PHILANTHROPY (Nov. 2022), <https://www.openphilanthropy.org/grants/alignment-research-center-general-support-november-2022/>.

Open Philanthropy, Center for Security and Emerging Technology — General Support, OPEN PHILANTHROPY (Aug. 2021), <https://www.openphilanthropy.org/grants/center-for-security-and-emerging-technology-general-support-august-2021/>.

Open Philanthropy, Georgetown University — Center for Security and Emerging Technology, OPEN PHILANTHROPY (Jan. 2019), <https://www.openphilanthropy.org/grants/georgetown-university-center-for-security-and-emerging-technology/>.

Open Philanthropy, RAND Corporation — Emerging Technology Fellowships and Research, OPEN PHILANTHROPY (Apr. 2023), <https://www.openphilanthropy.org/grants/rand-corporation-emerging-technology-fellowships-and-research/>.

Open Philanthropy, RAND Corporation — Emerging Technology Initiatives, OPEN PHILANTHROPY (July 2024), <https://www.openphilanthropy.org/grants/rand-corporation-emerging-technology-initiatives/>.

Open Philanthropy, RAND Corporation — Emerging Technology Initiatives, OPEN PHILANTHROPY (2025), <https://www.openphilanthropy.org/grants/rand-corporation-emerging-technology-initiatives-2025/>.

Open Philanthropy, RAND Corporation — Technology and Security Policy Center, OPEN PHILANTHROPY (Oct. 2024), <https://www.openphilanthropy.org/grants/rand-corporation-technology-and-security-policy-center/>.

P

Perrigo, Billy, How Anthropic Designed Itself to Avoid OpenAI's Mistakes, TIME (May 20, 2024), <https://time.com/6983420/anthropic-structure-openai-incentives/>.

Philanthropy News Digest, Open Philanthropy Project Awards \$55 Million for Center on AI Policy, PHILANTHROPY NEWS DIGEST (Feb. 28, 2019), <https://philanthropynewsdigest.org/news/open-philanthropy-project-awards-55-million-for-center-on-ai-policy>.

R

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

S

Salvaggio, Eryk, The AI Safety Debate Needs AI Skeptics, TECHPOLICY.PRESS, <https://www.techpolicy.press/the-ai-safety-debate-needs-ai-skeptics/>.

T

Tiku, Nitasha, Anthropic, OpenAI's Biggest Rival, Has a Weird Structure Too, SLATE (Dec. 4, 2023), <https://slate.com/technology/2023/12/anthropic-openai-board-trust-effective-altruism.html>.

U

U.K. Government, Tackling AI Security Risks to Unleash Growth and Deliver Plan for Change, GOV.UK (Feb. 14, 2025), <https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change>.

U.S. Department of Commerce, U.S. Commerce Secretary Gina Raimondo Announces Expansion of U.S. AI Safety Institute Leadership Team, U.S. DEP'T OF COM. (Apr. 16, 2024), <https://www.commerce.gov/news/press-releases/2024/04/us-commerce-secretary-gina-raimondo-announces-expansion-us-ai-safety>.

U.S. House of Representatives, Artificial Intelligence: Advancing Innovation Towards the National Interest, H. Comm. on Sci., Space, & Tech. (June 22, 2023), <https://science.house.gov/2023/6/artificial-intelligence-advancing-innovation-towards-the-national-interest>.

W

Wiggers, Kyle, A Timeline of Sam Altman's Firing from OpenAI -- and the Fallout, TECHCRUNCH (Jan. 5, 2024), <https://techcrunch.com/2024/01/05/a-timeline-of-sam-altmans-firing-from-openai-and-the-fallout/>.

Wiggers, Kyle, NIST Staffers Revolt Against Expected Appointment of 'Effective Altruist' AI Researcher to US AI Safety Institute, VENTUREBEAT (Mar. 8, 2024), <https://venturebeat.com/ai/nist-staffers-revolt-against-potential-appointment->

of-effective-altruist-ai-researcher-to-us-ai-safety-institute/.

Wiggers, Kyle, OpenAI and Anthropic Agree to Send Models to US Government for Safety Evaluations, VENTUREBEAT (Aug. 29, 2024), <https://venturebeat.com/ai/openai-and-anthropic-agree-to-send-models-to-us-government-for-safety-evaluations/>.

Additional Sources by Topic

Claude Model Technical Specifications

Anthropic, Model Card: Claude 2 (July 2023), <https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf>.

Anthropic, Model Card: Claude 3 Addendum (June 2024), https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Financial Reporting

CNBC, Amazon to Invest Another \$4 Billion in Anthropic, OpenAI's Biggest Rival, CNBC (Nov. 22, 2024), <https://www.cnbc.com/2024/11/22/amazon-to-invest-another-4-billion-in-anthropic-openais-biggest-rival.html>.

CNBC, Amazon Spends \$2.75 Billion on AI Startup Anthropic in Its Largest Venture Investment Yet, CNBC (Mar. 27, 2024), <https://www.cnbc.com/2024/03/27/amazon-spends-2point7b-on-startup-anthropic-in-largest-venture-investment.html>.

CNBC, Anthropic is Lining Up a New Slate of Investors, But the AI Startup Has Ruled Out Saudi Arabia, CNBC (Mar. 22, 2024), <https://www.cnbc.com/2024/03/22/anthropic-lining-up-a-new-slate-of-investors-ruled-out-saudi-arabia.html>.

CNBC, Google Commits to Invest \$2 Billion in OpenAI Competitor Anthropic, CNBC (Oct. 27, 2023), <https://www.cnbc.com/2023/10/27/google-commits-to-invest-2-billion-in-openai-competitor-anthropic.html>.

Decrypt, US Prosecutors Call SBF's \$500 Million Investment in AI Firm Anthropic 'Wholly Irrelevant,' DECRYPT (Oct. 8, 2023), <https://decrypt.co/200649/u-s-prosecutors-say-sbf-500-million-investment-ai-firm-anthropic-wholly-irrelevant>.

Fortune, Prosecutors Don't Want Bankman-Fried's Lawyers Mentioning His \$500 Million Investment in AI Startup Anthropic, FORTUNE (Oct. 9, 2023), <https://fortune.com/crypto/2023/10/09/sam-bankman-fried-trial-500-million-anthropic-investment-ai-claude/>.

TechCrunch, AI's Proxy War Heats Up as Google Reportedly Backs Anthropic with \$2B, TECHCRUNCH (Oct. 27, 2023), <https://techcrunch.com/2023/10/27/ais-proxy-war-heats-up-as-google-reportedly-backs-anthropic-with-2b/>.

TechCrunch, Amazon to Invest Up to \$4 Billion in AI Startup Anthropic, TECHCRUNCH (Sept. 25, 2023), <https://techcrunch.com/2023/09/25/amazon-to-invest-up-to-4-billion-in-ai-startup-anthropic/>.

TechCrunch, Anthropic Raises **13B Series F** at 183B Valuation, TECHCRUNCH (Sept. 2, 2025), <https://techcrunch.com/2025/09/02/anthropic-raises-13b-series-f-at-183b-valuation/>.

TechCrunch, Netflix Co-Founder Reed Hastings Joins Anthropic's Board, TECHCRUNCH (May 28, 2025), <https://techcrunch.com/2025/05/28/netflix-co-founder-reed-hastings-joins-anthropics-board/>.

Government Sources

FedScoop, Jason Matheny Named IARPA Director, FEDSCOOP (2015), <https://fedscoop.com/jason-matheny-named-iarpa-director/>.

FedScoop, 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/>.

IARPA, History, INTELLIGENCE ADVANCED RSCH. PROJECTS ACTIVITY, <https://www.iarpa.gov/who-we-are/history>.

Wikipedia and Reference Sources

Wikipedia, Alignment Research Center, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/Alignment_Research_Center.

Wikipedia, Anthropic, WIKIPEDIA (last visited Oct. 24, 2025), <https://en.wikipedia.org/wiki/Anthropic>.

Wikipedia, Center for Security and Emerging Technology, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/Center_for_Security_and_Emerging_Technology.

Wikipedia, Daniela Amodei, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/Daniela_Amodei.

Wikipedia, Holden Karnofsky, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/Holden_Karnofsky.

Wikipedia, Jason Gaverick Matheny, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/Jason_Gaverick_Matheny.

Wikipedia, National Security Commission on Artificial Intelligence, WIKIPEDIA (last visited Oct. 24, 2025), https://en.wikipedia.org/wiki/National_Security_Commission_on_Artificial_Intelligence.

Wikipedia, Paul Christiano (Researcher), WIKIPEDIA (last visited Oct. 24, 2025), [https://en.wikipedia.org/wiki/Paul_Christiano_\(researcher\)](https://en.wikipedia.org/wiki/Paul_Christiano_(researcher)).

Notes on Citation Quality and Verification

Primary Sources Used:

- Official Anthropic blog posts and announcements
- U.S. government sources (NIST, Department of Commerce, IARPA, Congressional testimony)
- U.K. government sources (GOV.UK)
- Academic papers (ArXiv, NeurIPS proceedings)
- Delaware Code statutes
- Harvard Law School Forum on Corporate Governance
- Open Philanthropy grants database

- Company press releases (Amazon, RAND Corporation, Georgetown University)

Secondary Sources Used:

- Major news outlets (Washington Post, New York Times, CNN, NPR)
- Tech journalism (TechCrunch, VentureBeat, CNBC, Fortune)
- AI research forums (AI Alignment Forum, LessWrong, Effective Altruism Forum)

Unverified Claims:

- Claude 1 parameter count of "~430 million" - could not be verified; estimates suggest ~175 billion parameters
- NBC News article on "anti-corruption infrastructure destruction" - not located despite extensive searches

Date Corrections:

- Claude 4 Opus released August 5, 2025 (ASL-3 protections activated May 2025)
- Constitutional AI paper published December 15, 2022
- Long-Term Benefit Trust announced October 2023

All URLs were verified as accessible as of October 24, 2025. This reference list provides comprehensive documentation for all verifiable claims in the research report.