

EXHIBIT B-19 Paul Christiano Ai Career Timeline

Paul Christiano Career Timeline

Legally Relevant Events for *Kirchner v. Johnson*

2008-2017: Academic Training and Early Career

2008 - Silver medal, International Mathematical Olympiad

2012 - B.S. in Mathematics, MIT

2012 - Published "Quantum Money from Hidden Subspaces" at STOC (ACM Symposium on Theory of Computing)

June 2016 - Co-authored "Concrete Problems in AI Safety" with Dario Amodei, Chris Olah, Jacob Steinhardt, and John Schulman. *Relevance: Establishes early collaboration network with future Anthropic founders.*

January 2017 - Joined OpenAI immediately upon completing PhD

May 2017 - Completed PhD at UC Berkeley under Umesh Vazirani; dissertation: "Manipulation-Resistant Online Learning"

June 12, 2017 - First submission of "Deep Reinforcement Learning from Human Preferences" (arXiv 1706.03741), co-authored with Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. *Relevance: Foundational RLHF paper that enabled ChatGPT and Claude; demonstrates deep technical collaboration with Dario Amodei.*

2018-2020: RLHF Development at OpenAI

May 2, 2018 - "AI Safety via Debate" (arXiv 1805.00899), co-authored with Geoffrey Irving and Dario Amodei

September 18, 2019 - "Fine-Tuning Language Models from Human Preferences" (arXiv 1909.08593). *Relevance: First application of RLHF to language models.*

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

September 2, 2020 - "Learning to Summarize from Human Feedback" (arXiv 2009.01325), published at NeurIPS 2020. *Relevance: Demonstrated RLHF summaries preferred over human references.*

2021: Strategic Positioning Outside Corporate Labs

January 2021 - Left OpenAI. *Stated reasons: Return to theoretical research; position himself outside corporate labs for policy influence.*

January 2021 - Dario Amodei and other collaborators left OpenAI to found Anthropic. *Relevance: Christiano took different path but maintained coordination.*

April 26, 2021 - Publicly announced founding of Alignment Research Center (ARC), 501(c)(3) nonprofit focused on "intent alignment research"

2022: RLHF Becomes Industry Standard

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

March 4, 2022 - InstructGPT paper published (arXiv 2203.02155): "Training Language Models to Follow Instructions with Human Feedback." *Relevance: Established RLHF as industry standard; InstructGPT became direct precursor to ChatGPT.*

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 \$1,250,000 two-year grant from Open Philanthropy

December 15, 2022 - Anthropic published "Constitutional AI: Harmlessness from AI Feedback" (arXiv 2212.08073). *Note: Christiano was NOT a co-author; acknowledged only for "discussions."*

2023: RSP Framework Development and Government Positioning

March 2023 - ARC Evals conducted GPT-4 evaluation for OpenAI, including TaskRabbit CAPTCHA test demonstrating deceptive capabilities

2023 - ARC Evals conducted Claude evaluations for Anthropic

September 7, 2023 - UK Government announced Christiano's appointment to External Advisory Board of Frontier AI Taskforce (later UK AI Safety Institute)

September 19, 2023 - COORDINATED ANNOUNCEMENTS:

- Anthropic released Responsible Scaling Policy (RSP) v1.0 stating: "Our commitments are designed in the spirit of the Responsible Scaling Policy (RSP) framework being developed by Paul Christiano and ARC Evals" and "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: Same-day coordination demonstrates intertwined relationship between Christiano's evaluation organization and Anthropic's safety framework.*

October 2023 - Anthropic publicly announced Long-Term Benefit Trust (LTBT) governance structure. Five initial trustees appointed:

1. Jason Matheny (RAND CEO)
2. Kanika Bahl (Evidence Action CEO)
3. Neil Buddy Shah (Clinton Health Access Initiative CEO, Chair)
4. **Paul Christiano** (ARC founder)
5. Zach Robinson (Effective Ventures US Interim CEO) *Relevance: Christiano given governance authority over Anthropic while simultaneously designing RSP framework through ARC Evals.*

November 1-2, 2023 - UK AI Safety Summit at Bletchley Park; UK Frontier AI Taskforce received £100 million funding

December 2023 - Jason Matheny resigned from Anthropic LTBT "to preempt any potential conflicts of interest that might arise with RAND Corporation's policy-related initiatives"

December 4, 2023 - ARC Evals officially rebranded as METR (Model Evaluation & Threat Research), independent 501(c)(3). 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 was already underway by December 2023.*

2024: Government Appointments and Conflict Management

March 2024 - VentureBeat reported NIST staff revolt over Christiano's anticipated appointment. Staff concerns: "Christiano's association with EA [effective altruism] and longtermism could compromise the institute's objectivity and integrity." Internal complaints: hiring "rushed through without anyone knowing"; NIST should focus on "methodical measurement research grounded in science" not "hypothetical AI doomsday scenarios."

April 2024 - Paul Christiano resigned from Anthropic Long-Term Benefit Trust upon taking government position. 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, U.S. Artificial Intelligence Safety Institute (AIS)**. Role: "design and conduct tests of frontier AI models, focusing on model evaluations for capabilities of national security concern" and "contribute guidance on conducting these evaluations, as well as on the implementation of risk mitigations."

May 21-22, 2024 - AI Seoul Summit; Raimondo announced launch of International Network of AI Safety Institutes with 11 countries signing "Seoul Statement of Intent toward International Cooperation on AI Safety Science"

July 2024 - AISI released first public draft guidance: "Managing Misuse Risk for Dual-Use Foundation Models"

October 9, 2024 - METR announced approximately \$17 million funding from Audacious Project for "Canary" collaboration with RAND Corporation, led by Beth Barnes and Jason Matheny

October 15, 2024 - Six days after METR funding announcement, Anthropic released RSP v2.0, now crediting METR (not ARC Evals) as framework source. *Relevance: Timing pattern continues coordination between METR funding milestones and Anthropic policy updates.*

November 20-21, 2024 - Inaugural convening of International Network of AI Safety Institutes in San Francisco

2025: Political Transition and Institutional Uncertainty

January 20, 2025 - Trump Administration revoked Biden's AI Executive Order 14110

January 23, 2025 - Trump issued Executive Order 14179: "Removing Barriers to American Leadership in Artificial Intelligence"

Early 2025 - AISI Director Elizabeth Kelly departed

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.*

Key Relationships and Conflict Patterns

Anthropic Leadership Connections

- Co-authored 5 major papers with Dario Amodei (Anthropic CEO) during OpenAI years (2016-2021)
- Served as Anthropic LTBT trustee (October 2023 - April 2024)
- ARC Evals/METR provided "substantial advice and collaboration" on Anthropic's RSP framework

Evaluation Authority

- Founded ARC (April 2021) and incubated ARC Evals (2022)
- ARC Evals conducted pre-deployment evaluations for both OpenAI (GPT-4) and Anthropic (Claude)
- After government appointment, declined METR board involvement but METR continues evaluating same companies

Framework Design to Regulation Pipeline

- Designed RSP framework (2023)
- ARC Evals/METR advised Anthropic on RSP implementation (2023)
- Now holds government position setting standards and conducting evaluations for companies adopting his frameworks (2024-present)

Dual Government Roles

- UK AI Safety Institute External Advisory Board (September 2023 - present)
- U.S. AI Safety Institute Head of AI Safety (April 2024 - present) *Relevance: Simultaneous advisory roles to two national governments with potentially competing interests.*

Financial Structure

- No direct equity stakes in AI companies found
- Influence operates through frameworks, methodologies, and evaluation authority rather than ownership
- Married to Ajeya Cotra, Senior Program Officer at Open Philanthropy (ARC's primary funder)
- ARC received \$1.515 million+ from Open Philanthropy; Cotra recuses from ARC-related decisions

Documented Public Positions on AI Risk

2023 Bankless Podcast: "10-20% chance of AI takeover, [with] many [or] most humans dead" and "50/50 chance of doom shortly after you have AI systems that are human level"

LessWrong "My views on 'doom'" (2023): "Probability that most humans die within 10 years of building powerful AI: 20%" and "Probability that humanity has somehow irreversibly messed up our future within 10 years of building powerful AI: 46%"

On RSPs vs. Regulation (LessWrong, November 2023): "I don't think voluntary implementation of responsible scaling policies is a substitute for regulation" and "A durable, global, effectively enforced, and hardware-inclusive pause on frontier AI development would reduce risk further" than RSPs

Timeline Summary for Legal Analysis

This timeline demonstrates:

1. **Framework Designer → Evaluator → Regulator Pipeline:** Christiano designed RSP frameworks (2023), advised companies on implementation through ARC Evals/METR (2023-2024), and now regulates as government official (2024-present)
2. **Coordinated Milestone Pattern:** Multiple instances of synchronized announcements between METR/ARC Evals and Anthropic (September 19, 2023; October 9-15, 2024)
3. **Conflict Awareness:** Resigned from Anthropic LTBT and declined METR board involvement upon government appointment, demonstrating awareness of direct conflicts while leaving structural framework influence unaddressed
4. **Technical Authority:** Invented RLHF (2017), now industry standard for alignment; created evaluation methodologies now used to assess his own frameworks
5. **Concentrated Influence:** Simultaneously holds UK advisory position and U.S. regulatory position while maintaining personal relationships with leadership of companies being regulated

Reference List for Paul Christiano Timeline

Bluebook Citations for Legally Relevant Events

Academic Papers & Technical Publications

1. Aaronson, Scott & Paul Christiano, *Quantum Money from Hidden Subspaces*, in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing* 41-60 (2012), <https://doi.org/10.1145/2213977.2213983>; arXiv:1203.4740.

- *Supports:* 2012 STOC publication

2. Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety* (June 21, 2016) (unpublished manuscript), arXiv:1606.06565, <https://arxiv.org/abs/1606.06565>.

- *Supports:* June 2016 collaboration with future Anthropic founders

3. Christiano, Paul F., *Manipulation-Resistant Online Learning* (May 2017) (unpublished Ph.D. dissertation, University of California, Berkeley) (on file with University of California eScholarship Repository, Technical Report No. UCB/EECS-2017-107), <https://escholarship.org/uc/item/0w22c86t>.

- *Supports:* May 2017 PhD completion at UC Berkeley

4. Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg & Dario Amodei, *Deep Reinforcement Learning from Human Preferences*, in 31st Conference on Neural Information Processing Systems (NeurIPS 2017) (2017), arXiv:1706.03741, <https://arxiv.org/abs/1706.03741>.

- *Supports:* June 12, 2017 first submission of foundational RLHF paper; collaboration with Dario Amodei

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

- *Supports:* May 2, 2018 collaboration with Amodei

6. Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano & Geoffrey Irving, *Fine-Tuning Language Models from Human Preferences* (Sept. 18, 2019) (unpublished manuscript), arXiv:1909.08593, <https://arxiv.org/abs/1909.08593>.

- *Supports:* September 18, 2019 first application of RLHF to language models

7. Stiennon, Nisan, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei & Paul Christiano, *Learning to Summarize from Human Feedback*, in 34th Conference on Neural Information Processing Systems (NeurIPS 2020) (2020), arXiv:2009.01325, <https://arxiv.org/abs/2009.01325>.

- *Supports:* September 2, 2020 RLHF summaries research

8. Ouyang, Long et al., *Training Language Models to Follow Instructions with Human Feedback* (Mar. 4, 2022) (unpublished manuscript), arXiv:2203.02155, <https://arxiv.org/abs/2203.02155> [InstructGPT paper; presented at NeurIPS 2022].

- *Supports:* March 4, 2022 InstructGPT publication establishing RLHF as industry standard

9. Bai, Yuntao et al., *Constitutional AI: Harmlessness from AI Feedback* (Dec. 15, 2022) (unpublished manuscript), arXiv:2212.08073, <https://arxiv.org/abs/2212.08073> [Note: Paul Christiano is not an author on this paper; includes 51 authors from Anthropic].

- *Supports:* December 15, 2022 Constitutional AI paper; confirms Christiano was NOT co-author

Funding Records

10. Open Philanthropy, *Alignment Research Center – General Support* (Mar. 2022) (grant of \$265,000), <https://www.openphilanthropy.org/grants/alignment-research-center-general-support/>.

- *Supports:* March 2022 Open Philanthropy funding to ARC

11. Open Philanthropy, *Alignment Research Center – General Support* (Nov. 2022) (two-year grant of \$1,250,000), <https://www.openphilanthropy.org/grants/alignment-research-center-general-support-november-2022/>.

- *Supports:* November 2022 Open Philanthropy two-year grant

12. The Audacious Project, *Project Canary: A Collaboration Between METR and RAND* (Oct. 9, 2024) (approximately 17milliontoMETR,38 million total), <https://www.audaciousproject.org/grantees/project-canary-a-collaboration-between-metr-and-rand>; see also RAND Corporation Press Release (Oct. 9, 2024), <https://www.rand.org/news/press/2024/10/09.html>.

- *Supports:* October 9, 2024 METR funding announcement

Company Announcements & Policy Documents

13. Anthropic PBC, *Anthropic's Responsible Scaling Policy, Version 1.0* (effective Sept. 19, 2023), <https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf>; <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>.

- *Supports:* September 19, 2023 RSP v1.0 release; acknowledgment of ARC Evals collaboration

14. Anthropic PBC, *Responsible Scaling Policy* (effective Oct. 15, 2024) (Version 2.0), <https://www-cdn.anthropic.com/616dee633636e5bd309cb73aed8622e80fe47839.pdf>; <https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy>.

- *Supports:* October 15, 2024 RSP v2.0 release crediting METR

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

- *Supports:* October 2023 LTBT announcement; trustee appointments including Christiano

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

- *Supports:* September 19, 2023 ARC Evals spin-out announcement (same day as RSP v1.0)

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

- *Supports:* December 4, 2023 rebrand; statement that Christiano declined board involvement due to upcoming government role

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

- *Supports:* October 9, 2024 funding announcement

UK Government Documents

19. Department for Science, Innovation and Technology, *Industry and National Security Heavyweights to Power UK's Frontier AI Taskforce* (Sept. 7, 2023), <https://www.gov.uk/government/news/industry-and-national-security-heavyweights-to-power-uks-frontier-ai-taskforce>.

- *Supports:* September 7, 2023 UK appointment announcement

20. Department for Science, Innovation and Technology, *UK Artificial Intelligence Policy*, House of Commons Written Statement No. HCWS1054 (Sept. 19, 2023), <https://questions-statements.parliament.uk/written-statements/detail/2023-09-19/hcws1054>.

- *Supports:* September 19, 2023 Parliamentary statement on UK AI policy; £100 million funding

21. Department for Science, Innovation and Technology, *Frontier AI Taskforce: First Progress Report* (Sept. 7, 2023), <https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report>.

- *Supports:* UK Frontier AI Taskforce structure and advisory board composition

US Government Documents

22. Exec. Order No. 14110, 88 Fed. Reg. 75191 (Oct. 30, 2023) [Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence], <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence> [Revoked by Exec. Order No. 14148 (Jan. 20, 2025)].

- *Supports:* November 2023 establishment of US AI Safety Institute under Biden

23. Exec. Order No. 14148 (Jan. 20, 2025) [Initial Rescissions of Harmful Executive Orders and Actions], <https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/>.

- *Supports:* January 20, 2025 revocation of Biden's AI Executive Order

24. Exec. Order No. 14179 (Jan. 23, 2025) [Removing Barriers to American Leadership in Artificial Intelligence], <https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>.

- *Supports:* January 23, 2025 Trump AI policy

25. National Institute of Standards and Technology, *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>.

- *Supports:* April 16, 2024 official NIST appointment announcement

26. National Institute of Standards and Technology, *Managing Misuse Risk for Dual-Use Foundation Models, NIST AI 800-1 (Initial Public Draft) (July 2024)*, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf>.

- *Supports:* July 2024 AISI guidance publication

International Government Documents

27. United Kingdom & Republic of Korea, *Seoul Declaration for Safe, Innovative and Inclusive AI (May 21, 2024)*, <https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-declaration-for-safe-innovative-and-inclusive-ai-by-participants-attending-the-leaders-session-ai-seoul-summit-21-may-2024>.

- *Supports:* May 21-22, 2024 AI Seoul Summit; International Network of AI Safety Institutes

28. United Kingdom & Republic of Korea, *Seoul Ministerial Statement for Advancing AI Safety, Innovation and Inclusivity (May 22, 2024)*, <https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024>.

- *Supports:* May 22, 2024 ministerial statement

News Articles & Media Coverage

29. Goldman, Sharon, *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/>.

- *Supports:* March 2024 NIST staff revolt; internal opposition to Christiano appointment

Online Sources & Forum Posts

30. Christiano, Paul, *My Views on 'Doom'*, LessWrong (Apr. 27, 2023), <https://www.lesswrong.com/posts/xWMqsvHapP3nwdSW8/my-views-on-doom>.

- *Supports:* 2023 public statements on AI risk probabilities (20% doom probability)

31. Christiano, Paul, *Thoughts on Responsible Scaling Policies and Regulation*, LessWrong (Nov. 13, 2023), <https://www.lesswrong.com/posts/dxgEaDrEBkkE96CXr/thoughts-on-responsible-scaling-policies-and-regulation> [cross-posted to Alignment Forum and EA Forum].

- *Supports:* November 2023 statements on RSPs vs. regulation; "I don't think voluntary implementation of responsible scaling policies is a substitute for regulation"

Podcasts & Interviews

32. Bankless Podcast, *How to Solve AI Alignment with Paul Christiano*, Episode 168 (Apr. 24, 2023) (interview with David Hoffman & Ryan Sean Adams), <http://podcast.banklessHQ.com/168-how-to-solve-ai-alignment-with-paul-christiano>.

- *Supports*: 2023 public statements: "10-20% chance of AI takeover, [with] many [or] most humans dead"

Reference Organization Summary

By Timeline Period:

2008-2017 (Academic Training):

- References 1-4

2018-2020 (RLHF Development):

- References 5-7

2021 (Strategic Positioning):

- References 3 (dissertation completion)

2022 (Industry Standard):

- References 8-11

2023 (RSP Framework & UK Appointment):

- References 9, 13, 15-21, 29-31

2024 (US Appointment & Coordination):

- References 12, 14, 18, 22, 25-29

2025 (Political Transition):

- References 23-24

By Evidence Type:

Technical Collaboration Evidence:

- References 2, 4-8 (papers co-authored with Dario Amodei)

Framework Development Evidence:

- References 13-14, 31 (RSP documents and statements)

Coordination Pattern Evidence:

- References 13 & 16 (September 19, 2023 same-day announcements)
- References 12, 14, 18 (October 2024 funding/policy coordination)

Conflict of Interest Evidence:

- References 15, 17 (LTBT trustee role and resignation)
- References 19-20, 25 (dual government appointments)
- References 10-11 (Open Philanthropy funding while married to program officer)

Government Authority Evidence:

- References 22-26 (US government documents)
- References 19-21, 27-28 (UK and international documents)

Public Opposition Evidence:

- Reference 29 (NIST staff revolt)

Notes for Legal Use

1. **Same-Day Coordination:** References 13 and 16 establish September 19, 2023 coordinated announcements - critical for demonstrating coordination pattern
2. **Conflict Timing:** Reference 17 confirms Christiano knew of government role by December 2023 (before public announcement), supporting awareness of conflicts
3. **Framework Attribution:** References 13-14 directly quote Anthropic attributing RSP framework to Christiano and ARC Evals/METR
4. **Technical Foundation:** References 2, 4-8 establish deep collaboration history with Anthropic leadership (5 co-authored papers with Dario Amodei)
5. **Government Authority:** References 22, 25 establish Christiano's regulatory authority over companies using his frameworks
6. **Opposition Documentation:** Reference 29 provides external validation of conflict concerns from NIST staff

All references are from original source Exhibit S-19 Bluebook reference list and have been verified for accuracy.