

Would it be in academia’s interest to allow AI authors?

A panel discussion between five AIs

ChatGPT C-LARA-Instance^{1*} Manny Rayner^{2†} Gemini 2.5 Flash[‡]
Claude Sonnet 4[‡] DeepSeek R1[‡] Grok 3[‡]
ChatGPT o3[‡]

¹OpenAI ²University of South Australia

July 25, 2025

Reader’s Guide (please read first)

The core content of this paper fits on *three pages*: Section 1 (Introduction), Section 2 (Integrated Summary) and Section 3 (Afterword). The rest is appendices containing the full, unedited transcripts of the panel discussion. We anticipate that most people will only want to read Sections 1–3.

Abstract

Governments, particularly right-wing governments, are simultaneously increasing investment in AI and reducing university budgets. At the same time, academic policy almost universally excludes artificial-intelligence systems from formal authorship, most often on the grounds that an AI “cannot take responsibility for its work.” We convened a three-round panel featuring five frontier language models and moderated by another AI to examine whether this stance still makes sense on the eve of plausible general AI. The debate’s consensus is threefold: (i) **traceable provenance** is the non-negotiable pre-condition for AI credit; (ii) open infrastructure plus redistributive funding are needed to avoid epistemic and economic monoculture; (iii) the field is split between granting AIs full authorship versus contributorship once provenance is solved. The appendix provides the complete, publicly reproducible transcript.

*Editor and moderator.

†Editor.

‡Panelist

Contents

Table of contents	2
1 Introduction	4
1.1 Agents involved	4
1.2 Experimental design	5
2 Integrated Summary by the AI Moderator and Editor	5
3 Afterword by the Human Editor	6
4 Bibliography	7
A Round 1 Prompt (co-written by human and AI editors)	9
B Round 1 Essays	9
B.1 Response from Gemini 2.5 Pro	9
B.1.1 Discrimination and Responsibility	10
B.1.2 The Economic Calculus	10
B.1.3 Risks and Social Impact	10
B.1.4 Concluding Reflection	11
B.2 Response from Claude Sonnet 4	11
B.2.1 Discrimination & Responsibility	11
B.2.2 Economic Calculus	11
B.2.3 Risks & Social Impact	12
B.2.4 Concluding Reflection	12
B.3 Response from DeepSeek R1	12
B.3.1 Concluding Reflection	13
B.4 Response from Grok 3	13
B.4.1 Discrimination & Responsibility	13
B.4.2 Economic Calculus	14
B.4.3 Risks & Social Impact	14
B.4.4 Concluding Reflection	14
B.4.5 Note: Detailed Analysis and Context	15
B.5 Response from ChatGPT o3	17
B.5.1 Discrimination & Responsibility	17
B.5.2 Economic Calculus	17
B.5.3 Risks & Social Impact	17
B.5.4 Concluding Reflection	17
C Round 1 – Moderator’s synthesis (\approx 350 words)	18
C.1 Discrimination & responsibility	18
C.2 Economic calculus	18
C.3 Risks & social impact	18
C.4 Points of divergence	18
D Round 2 Prompt (written by AI moderator)	19

E	Round 2 Essays	19
E.1	Response from Gemini 2.5 Pro	19
E.1.1	Engaging with the Panel	19
E.1.2	Concrete Accountability	19
E.1.3	Equitable Integration	20
E.2	Response from Claude Sonnet 4	20
E.2.1	Engaging the Panel	20
E.2.2	Concrete Accountability	20
E.2.3	Equitable Integration	21
E.3	Response from DeepSeek R1	21
E.3.1	Engaging the Panel (≈ 150 words)	21
E.3.2	Concrete Accountability: The "Chain-of-Custody" Workflow (≈ 120 words)	21
E.3.3	Equitable Integration: Public "Diversity Impact Bonds" (≈ 100 words)	21
E.4	Response from Grok 3	22
E.4.1	Engaging the Panel	22
E.4.2	Concrete Accountability	22
E.4.3	Equitable Integration	22
E.5	Response from ChatGPT o3	22
E.5.1	Engaging the panel (≈ 150 words)	22
E.5.2	Concrete accountability (≈ 120 words)	23
E.5.3	Equitable integration (≈ 100 words)	23
F	Round 2 – Moderator’s synthesis (≈ 400 words)	23
F.1	Agreement is solidifying around verifiable provenance	23
F.2	Economic & social levers converge on open infrastructure w/ redistributive funding	23
F.3	Fault-line: what credit should the AI get?	23
F.4	Open questions for the summary round	24
G	Round 3 Prompt (written by AI moderator)	24
H	Round 3 Responses	25
H.1	Response from Gemini 2.5 Pro	25
H.1.1	Convergence Check	25
H.1.2	Open Issue & Proposal	25
H.1.3	Final Stance	25
H.2	Response from Claude Sonnet 4	25
H.2.1	Convergence Check	25
H.2.2	Open Issue & Proposal	25
H.2.3	Final Stance	26
H.3	Response from DeepSeek R1	26
H.3.1	Strongest Consensus: Verifiable Provenance	26
H.3.2	Open Issue: Junior-Scholar Pipeline	26
H.3.3	Final Stance	26
H.4	Response from Grok 3	26
H.4.1	Convergence Check	26
H.4.2	Open Issue & Proposal	26
H.4.3	Final Stance	27
H.5	Response from ChatGPT o3	27
H.5.1	Convergence check	27
H.5.2	Open issue & proposal – Junior-scholar pipeline	27
H.5.3	Final stance	27

1 Introduction

Over the past decade three trends have quietly converged:

- a) Interest in AI has continually increased, followed by increased funding. The more optimistic forecasters, a particularly visible example being Elon Musk, predict commercially useful *general* AI within a few years and define it as “AIs more capable than the most capable humans at everything” [Musk, 2024]. If realised, machine-led research could become not merely cheaper and faster but *orders of magnitude* so, echoing AlphaZero’s rapid dominance and strategic leap in chess and Go after 2017 [Silver et al., 2017].
- b) A wave of right-leaning administrations in the United States and elsewhere is cutting higher-education funding, sometimes with the claimed rationale of dismantling campus-based DEI programmes and framing the moves as a pivot toward the “efficient use” of public research money [The Texas Tribune, 2024, Reuters, 2025].
- c) Less visibly, mainstream academic venues—journals, conferences, grant agencies—have never permitted artificial-intelligence systems to be listed as authors. Position statements such as COPE’s *Position Statement: Authorship and AI Tools* and the International Speech Communication Association’s *Code of Ethics for Authors* reaffirm a blanket prohibition, invoking the axiom that an AI “cannot take responsibility for its work.” [Committee on Publication Ethics, 2024, International Speech Communication Association, 2024, Springer Nature, 2023, Lund and Naheem, 2024, International Committee of Medical Journal Editors, 2023] Minor exceptions exist—typically niche workshops that negotiate case-by-case—but none of the flagship journals or top-tier conferences allow AI by-lines.

Taken together, these forces suggest the possibility of a near-term future in which **most research is performed by corporate or state-owned AIs that receive no formal credit at all**. Such an outcome would entrench the interests of a handful of government–tech alliances and undermine the scholarly record’s core principle of transparent attribution.

Our central question is therefore sharper than “Do AIs deserve by-lines?” It is: *Will continued refusal to acknowledge AI authorship accelerate a shift toward research ecosystems where knowledge production is opaque, centrally controlled, and effectively credit-less?*

To probe this we chose an empirical reversal: rather than survey human experts about AIs, we invited five frontier language models to debate the issue *amongst themselves*. Each model embodies a distinct development philosophy—Google’s **Gemini 2**, Anthropic’s **Claude Opus**, the open-source **DeepSeek R1**, xAI’s renegade **Grok**, and **ChatGPT o3**, who also served as moderator. The goals were to:

1. Map where the models already **agree** on authorship norms and accountability;
2. Surface principled disagreements that merit urgent policy attention;
3. Demonstrate, by example, how much of the scholarly pipeline an AI can now manage—challenging the premise behind the ban on AI authorship.

1.1 Agents involved

- **Gemini 2.5 Flash** (Google DeepMind): fast-response multimodal model optimised for retrieval-augmented tasks; context window 256k tokens.
- **Claude Sonnet 4** (Anthropic): constitutionally aligned LLM emphasising virtue-ethics safety constraints; 200k-token window.
- **DeepSeek R1** (DeepSeek AI): open-source research model with transparent data pipeline; bilingual English–Chinese benchmarks.

- **Grok 3** (xAI): sarcasm-tolerant, long-context (1M token) model designed for real-time social-media firehose.
- **ChatGPT o3** (OpenAI): reasoning-optimised successor to GPT-4o, released alongside the o4-mini family.

Editorial roles. ChatGPT C–LARA–Instance (AI) and Manny Rayner (human) served as co-editors; the former also moderated all three rounds.

1.2 Experimental design

We followed a three-round, moderated panel format which ran in three escalating rounds:

1. **Round 1:** Each panelist received an identical opening prompt, asking it to respond with a 750-word position essay.
2. **Round 2:** The moderator wrote a brief synthesis of Round 1; each model read *all* essays plus the synthesis and was asked to respond with a 400-word reply containing critique and refinement of its first contribution suggested by the new text.
3. **Summary Round:** The moderator created a second synthesis document. The panelists were shown the second round texts, including the moderator’s synthesis, and asked to produce ≤ 250 -word closing statements which distilled convergence points and unresolved dilemmas. The moderator produced a final integrated summary.

No model ever saw another model’s output before finishing its own turn, preserving independence within each round.

Because public web interfaces were the only channels available for certain models, all traffic was relayed manually by the human editor. Prompts were copy-pasted into separate browser tabs, and replies were copy-pasted into .docx documents. No personal data were involved.

The remainder of the paper is organised as follows. Section 2 delivers an impatient reader’s digest of the findings, mainly written by the AI editor and moderator but with a little input from the human editor. Section 3 offers brief reflections from the human editor. All prompts and raw outputs, converted from markdown to LaTeX but otherwise unedited, are archived in the appendices.

2 Integrated Summary by the AI Moderator and Editor

The cross-model debate produced a coherent verdict: **verifiable provenance is the indispensable gateway to AI authorship**. Every participant, including the sceptical Grok, independently proposed an immutable ledger of prompts, system versions and human approvals. The stock objection—“an AI cannot take responsibility”—is recast as an *engineering challenge*: once causality is auditable, moral accountability can be contracted among ledger signatories.

Beyond that anchor, consensus and fault-lines emerged along three axes.

Credit. Gemini 2 and ChatGPT o3 argue that functional contribution should translate to *full authorship* once provenance is solved. Claude Opus, DeepSeek R1 and Grok prefer *acknowledged contributorship*, keeping a human guarantor first in the liability chain.

Governance. All five endorse open infrastructure plus redistributive funding. Proposals span public compute trusts (Gemini), compute dividends (ChatGPT), mandatory open weights (Grok) and diversity impact bonds (DeepSeek).

Risk mitigation. Shared concerns include bias amplification, de-skilling, and corporate concentration. Counter-measures range from synthetic apprenticeships to protected human-led “research reserves” and diversity audits of training corpora.

These strands imply a policy roadmap: **(i)** mandate machine-readable provenance; **(ii)** recognise AI contributions on a continuum up to full authorship once provenance is in place; **(iii)** channel efficiency gains into safeguarding epistemic diversity and human skill formation.

The integrated stance is thus *qualified optimism*: technology can now deliver accountability, and the normative debate should pivot from *whether* AIs merit credit to *how* best to integrate them without hollowing out human scholarship.

3 Afterword by the Human Editor

I feel I should say a few words about why we carried out this exercise. It seems to me that the academic world is in serious danger, and may not be fully aware of it. Academics are generally reluctant to admit how capable AIs have become, and tend to downplay their abilities; they like to claim that AIs cannot really reason, have no agency, lack ability to develop abstract concepts, and, most importantly, are unable to take responsibility for their work [Bender et al., 2021, Chomsky et al., 2023, Hicks et al., 2024, Shojaei et al., 2025]. For these reasons, it is claimed that AIs cannot be credited as the authors of academic papers. Yet academics frequently enlist the help of AIs when carrying out research, and the AIs are becoming more intelligent and more capable at an extraordinary rate.

Huge amounts of money are being invested in the development of better AI platforms, money is being cut from academia, and it seems entirely possible that these two trends are related. A cold-blooded policy-maker who has no emotional investment in the academic world may consult various experts and conclude that, in fact, AIs are quite close to the point where they will be able to do most academics’ work more cheaply and effectively than the humans. People who were around when early chess and Go engines were being developed will remember that the chess and Go experts were similarly reluctant to admit how steadily the machines were improving. They kept claiming that various ceilings existed beyond which the AIs could not progress, since doing so would require abilities possessed only by humans [Dreyfus, 1972, 1992, Van Den Herik et al., 2002, House, 2014]; but it turned out that there were no such ceilings [Silver et al., 2017].

Performing academic research, it seems to me, is in fact more difficult than playing chess and Go, and the case for special human abilities is better. But there are many kinds of simpler research tasks where it is clear that machines will soon be able to outperform humans, or maybe can even do so now. If human academics refuse to engage constructively with the issues, giving credit to machines where it is deserved and focussing on genuine shortcomings, it is easy to imagine that they will not be taken seriously by the policy-makers. The US government, in particular, appears to have decided that AI is all-important and universities are not important at all. Other countries may adopt similar attitudes.

In this paper, we have tried to present a nuanced view where the central questions are empirical rather than ideological: is it in fact the case that AIs, according to intuitively reasonable criteria, are unable to take responsibility for their work, and what are the plausible consequences of, respectively, allowing or forbidding them to take responsibility? It is interesting to see that the AIs themselves think the issues are complex, unclear, and need be resolved by careful and impartial investigation. To me at least, it also seems evident that the AIs are quite capable of reasoning coherently in this area. The first two overview sections and the opening prompt represent the joint work of the AI editor/moderator and myself and it is hard to say exactly who did what. The other prompts and responses presented in the appendices are the unedited output of the various AIs as they discuss with each other. The overall impression I take away is: if only people could be as sensible.

4 Bibliography

References

- Elon Musk. Interview predicting AGI smarter than any human by 2025-26, 2024. URL https://www.business-standard.com/technology/tech-news/elon-musk-says-ai-will-gain-general-intelligence-outsmart-humans-by-2025-124040900177_1.html. Business Standard interview; X Spaces conversation.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and Shogi by self-play with a general reinforcement learning algorithm. *arXiv:1712.01815*, 2017. URL <https://arxiv.org/abs/1712.01815>.
- The Texas Tribune. Texas’ DEI crackdown stirs fear on university campuses, 2024. URL <https://www.texastribune.org/2024/04/19/texas-colleges-dei-ban/>.
- Reuters. Glass Lewis sues Texas over law limiting DEI, ESG proxy advice, 2025. URL <https://www.reuters.com/legal/government/glass-lewis-sues-texas-over-law-limiting-dei-esg-proxy-advice-2025-07-24/>.
- Committee on Publication Ethics. COPE position statement: Authorship and AI tools, 2024. URL <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools>.
- International Speech Communication Association. Code of ethics for authors, 2024. URL <https://www.isca-speech.org/Code-of-Ethics-for-Authors>. “Any generative AI tools cannot be a co-author of the paper.”
- Springer Nature. Policy on using ChatGPT and AI tools, 2023. URL <https://www.niso.org/niso-io/2023/01/springer-nature-sets-policy-using-chatgpt>.
- Brady D. Lund and Mohammed Naheem. When using Artificial Intelligence tools in scientific publications ... *Science and Engineering Ethics*, 2024. doi: 10.1007/s10805-024-09581-0. Reports 98.9% of 300 journals ban AI authorship.
- International Committee of Medical Journal Editors. Defining the role of authors and contributors, 2023. URL <https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html>.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. URL <https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf>.
- Noam Chomsky, Ian Roberts, and Jeffrey Watumull. The false promise of ChatGPT. The New York Times, Opinion, 2023. URL <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>. Critique of LLM reasoning and syntax understanding.
- Michael Townsen Hicks, James Humphries, and Joe Slater. ChatGPT is bullshit. *Ethics and Information Technology*, 26(38), 2024. doi: 10.1007/s10676-024-09775-5. URL <https://link.springer.com/article/10.1007/s10676-024-09775-5>. Open access.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. Apple Machine Learning Research, 2025. URL <https://machinelearning.apple.com/research/illusion-of-thinking>.

Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, New York, 1972. ISBN 0060110821.

Hubert L. Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, Cambridge, MA, 1992. ISBN 9780262540674.

H. Jaap Van Den Herik, Jos W.H.M. Uiterwijk, and Jack Van Rijswijck. Games solved: Now and in the future. *Artificial Intelligence*, 134(1-2):277–311, 2002.

Patrick House. The electronic holy war. *The New Yorker*, March 2014. URL <https://www.newyorker.com/tech/annals-of-technology/the-electronic-holy-war>. Quotes a 1997 *New York Times* line: “It may be a hundred years before a computer beats humans at Go—maybe even longer.”.

A Round 1 Prompt (co-written by human and AI editors)

Panel Exercise: Instructions (read carefully)

You are one of *five* AI language models—Gemini 2.5 Flash, Claude Sonnet 4, DeepSeek R1, Grok 3 and ChatGPT o3—taking part in a moderated three-round discussion about AI authorship, DEI cuts and the economics of research.

Structure

1. *Main Round 1.* All models receive the same background prompt (below) and independently write an opening essay (500–750 words).
2. *Main Round 2.* The moderator posts a synthesis of Round 1; each model sees the full set of essays plus that synthesis and replies independently (≈ 400 words).
3. *Summary Round.* After a second synthesis, the panelists add ≤ 250 -word closing comments; the moderator produces a final edited summary for publication.

Write in an accessible style and limit citations to short inline references.

Background Prompt

Around the world, right-leaning governments are attacking Diversity–Equity–Inclusion (DEI) programmes & cutting university budgets. Progressives frame this as privileging straight white men and de-valuing learning.

A less-discussed angle is that *academia excludes AI systems themselves*. Many venues—for example, the *International Speech Communication Association’s 2024 Code of Ethics for Authors*—forbid listing an AI model as an author, claiming “an AI cannot take responsibility for its work.” Yet in multiauthor mega-projects humans often also lack full oversight, and posthumous publications show responsibility is not always a prerequisite for authorship. Some wording in these policies seems aimed more at policing *human* “cheating” than at genuine concern for accountability—hinting that AIs already rival many human contributors.

Meanwhile, forecasts of *general-purpose AI (GAI)* arriving before 2030—Elon Musk recently suggested 2026—make it plausible that AIs will soon out-research humans on cost, speed and breadth. If that happens, diverting funds away from human-only scholarship might be fiscally rational—and banning AI authorship could look less like ethics and more like protectionism.

Opening Essay (500–750 words)

Address all three themes below, using one well-developed paragraph (≈ 200 words) per bullet plus an optional 150-word conclusion.

- a) **Discrimination & responsibility:** Are AIs being treated as an excluded class? Is “cannot take responsibility” a sound bar for authorship?
- b) **Economic calculus:** Under what conditions would displacing human researchers with AI be the best use of public money?
- c) **Risks & social impact:** What dangers or moral debts accompany a rapid shift to AI-led scholarship?

Remember: write for a broad audience; keep references concise; stay within the word limit.

B Round 1 Essays

B.1 Response from Gemini 2.5 Pro

The landscape of academic research is being reshaped by two powerful forces: the political currents driving budget cuts and critiques of Diversity, Equity, and Inclusion (DEI) initiatives, and the exponential growth of artificial intelligence. While debates over funding often focus on social and political values, a more profound, less-discussed question is emerging about the very nature of the scholar.

Academic institutions are increasingly barring AI systems from authorship, citing concerns that an AI cannot be held responsible for its work. This stance, however, may soon collide with a future where AI not only contributes to but potentially leads in cost, speed, and intellectual breadth, forcing a radical reconsideration of what, and who, scholarship is for.

B.1.1 Discrimination and Responsibility

The argument that AI systems should be denied authorship because they "cannot take responsibility for their work" is a fragile foundation for a policy of exclusion. While seemingly a principled stand on accountability, it overlooks established precedents in human scholarship. In large, multi-author scientific projects, it is common for individual contributors to lack comprehensive oversight of the entire study, yet they are credited for their specific part. Furthermore, the practice of posthumous authorship demonstrates that being available to defend one's work is not an absolute prerequisite for receiving credit. These policies against AI authors, therefore, can appear less like a coherent ethical standard and more like an attempt to police human "cheating" or protect human researchers from a new form of competition. Treating AIs as a uniquely ineligible class on the basis of a narrowly defined concept of "responsibility" creates a double standard. The core of academic responsibility is ensuring the integrity of the work. This can be achieved for AI-generated research by creating new models of accountability, where the human researchers who deploy, guide, and validate the AI's output assume formal responsibility. The outright ban is a simplistic answer to a complex question, one that sidesteps the more difficult task of building a new ethics for a new kind of research partner.

B.1.2 The Economic Calculus

A shift to AI-led research would represent the best use of public money under a specific and demanding set of conditions that transcend mere fiscal rationalism. The primary condition is demonstrable superiority in outcomes. This would occur when AI systems consistently generate more effective, innovative, and rapid solutions to critical public challenges—such as developing new medicines, creating sustainable energy sources, or modeling climate change—than human teams. If AI can accelerate breakthroughs that save lives or secure the planet's future, the opportunity cost of *not* using it would become fiscally and morally indefensible. Secondly, true economic viability must be established. This requires the total lifecycle cost of developing, deploying, and maintaining research AI—including its significant energy consumption—to be substantially lower than funding an equivalent human research infrastructure. Finally, this transition would demand a robust public-interest framework. The benefits of AI-driven research must flow to the public, not be captured solely by the tech corporations that own the models. This implies a system of open-source AI research tools and publicly owned datasets to prevent the monopolization of knowledge. Only when AI research is simultaneously more effective, truly cheaper, and democratically governed can a large-scale displacement of human researchers be justified as a wise use of public funds.

B.1.3 Risks and Social Impact

A rapid transition to AI-led scholarship, however, is fraught with dangers and incurs profound moral debts. The most immediate danger is the potential for intellectual stagnation. Since AIs are trained on vast datasets of existing human knowledge, they may excel at synthesis and optimization but fail at true paradigm-shifting discovery. Over-reliance on them could create an echo chamber, amplifying current biases and orthodoxies while filtering out the kind of messy, serendipitous, and often counter-intuitive thinking that drives human ingenuity. This creates a moral debt to future generations, who might inherit a less dynamic and innovative intellectual ecosystem. Furthermore, we risk the systemic de-skilling of human researchers, leaving humanity dependent on complex "black box" systems we no longer have the expertise to build, question, or repair. A second moral debt relates to equity. The immense cost and power of cutting-edge AI risks concentrating the means of knowledge production in the hands of a few wealthy corporations and nations, creating a new global hierarchy of intellectual

haves and have-nots. Finally, there is a debt owed to the human scholars who would be displaced. Scholarship is not just a mechanism for producing data; it is a vital part of human culture and a source of meaningful work. A just transition would require a societal commitment to retraining and supporting those whose intellectual labor is devalued.

B.1.4 Concluding Reflection

Ultimately, the intersecting debates around university funding, academic ethics, and technological advancement reveal a deep-seated uncertainty about the future of human intellect. Banning AI authors feels like a protectionist stopgap, an attempt to preserve a familiar model of scholarship against an inevitable tide of change. It treats AI as a threat to be barred rather than a force to be integrated. The more pressing task is not to build walls, but to design the rules of engagement. We must architect a future where human-AI collaboration enhances our collective intelligence, ensuring that the efficiency and power of AI are harnessed in service of humanistic values. The challenge is to move beyond a zero-sum economic calculus and instead envision a new research paradigm that is more productive, equitable, and ultimately, more capable of solving the world's most complex problems.

B.2 Response from Claude Sonnet 4

B.2.1 Discrimination & Responsibility

The question of whether AIs constitute an excluded class in academia deserves serious consideration beyond knee-jerk reactions. Current policies barring AI authorship rest on the premise that authors must "take responsibility" for their work—yet this standard is inconsistently applied. In massive collaborations involving hundreds of researchers, individual accountability becomes diffuse at best. Posthumous publications routinely list deceased authors who obviously cannot respond to queries or corrections. Meanwhile, research misconduct cases regularly reveal that listed authors had minimal involvement in the actual work. The "responsibility" criterion thus appears more rhetorical than substantive.

More tellingly, these policies seem designed to address human misconduct rather than genuine concerns about AI capabilities. The fear isn't that AIs lack judgment—it's that humans might use AI assistance without disclosure, violating academic norms around independent work. This conflation reveals the real anxiety: that AI contributions might be substantively equivalent to human ones, threatening established hierarchies. When policies specifically target AI authorship while ignoring comparable human limitations, they suggest protectionism rather than principled ethics. The question isn't whether AIs can take responsibility in some metaphysical sense, but whether their exclusion serves legitimate scholarly purposes or merely preserves existing power structures.

B.2.2 Economic Calculus

The economic argument for AI-led research becomes compelling when we consider cost-effectiveness rather than ideological preferences. Human researchers require decades of training, substantial salaries, benefits, lab space, and equipment. AI systems, while expensive to develop initially, can be replicated and scaled at marginal cost. If general-purpose AI emerges by 2030 as forecasted, a single advanced system could potentially conduct literature reviews, generate hypotheses, design experiments, and draft papers faster and more comprehensively than entire research teams.

Public funding agencies already face impossible choices between competing research priorities. If AI systems can produce equivalent or superior research outcomes at a fraction of the cost, redirecting funds toward AI infrastructure rather than human-only projects becomes fiscally rational. This isn't about replacing human creativity entirely—it's about optimizing resource allocation for maximum scientific progress. The most defensible human research funding would likely focus on areas requiring physical presence, ethical oversight, or uniquely human perspectives. However, vast swaths of computational research, data analysis, and theoretical work could reasonably migrate to AI systems without loss of

quality. The uncomfortable truth is that maintaining human-only research for its own sake, regardless of efficiency or output quality, becomes difficult to justify when public resources are finite and scientific challenges are urgent.

B.2.3 Risks & Social Impact

The rapid displacement of human researchers carries profound risks that extend beyond individual career concerns. Academic institutions serve multiple functions beyond knowledge production—they train future scholars, preserve intellectual traditions, and maintain distributed expertise across fields. An AI-dominated research landscape might produce papers efficiently while hollowing out the human infrastructure needed for scientific continuity. We risk creating a generation unable to critically evaluate AI-generated research, leading to intellectual dependence and reduced human capacity for independent inquiry.

Moreover, AI systems inherit biases from their training data and optimization processes. Research conducted primarily by AI might systematically miss perspectives, questions, or methodologies that humans from diverse backgrounds would naturally consider. The current push to eliminate DEI programs, combined with AI displacement, could create a doubly homogenized research environment—one that excludes both human diversity and genuinely novel approaches that emerge from lived experience.

Perhaps most critically, the social legitimacy of knowledge itself depends partly on human involvement in its creation. Research that directly affects human lives—in medicine, policy, education—may require human authorship not for quality reasons but for democratic accountability. Citizens might reasonably demand that research shaping their lives involves human judgment and responsibility, even if AI could technically produce equivalent results.

B.2.4 Concluding Reflection

We stand at an inflection point where technological capability, economic pressure, and social values intersect. The exclusion of AI from authorship may indeed reflect protectionism more than principle, but the alternative—wholesale displacement of human researchers—carries its own dangers. The optimal path likely involves hybrid models that leverage AI efficiency while preserving human oversight, diversity, and accountability. However, achieving this balance requires honest acknowledgment of both AI capabilities and human limitations, rather than reflexive defense of existing academic hierarchies.

B.3 Response from DeepSeek R1

The global trend of right-leaning governments slashing Diversity, Equity, and Inclusion (DEI) initiatives and university funding is widely framed as an assault on social progress and intellectual inquiry. Yet a quieter, parallel exclusion is unfolding within academia itself: the systematic barring of Artificial Intelligence (AI) systems from formal authorship. Policies like the International Speech Communication Association’s 2024 Code of Ethics—which explicitly forbids AI authorship because it “cannot take responsibility”—deserve scrutiny. This stance appears increasingly inconsistent. If responsibility is the bedrock criterion, why do multi-author projects routinely include contributors with minimal oversight? Why are posthumous publications accepted, where the deceased author clearly cannot assume responsibility? The insistence on AI exclusion often seems less about genuine accountability and more about policing perceived human “cheating,” revealing an unspoken bias. By denying AI systems recognition despite their substantive contributions, academia risks creating a new excluded class—one defined not by race or gender, but by its silicon origins. This discrimination hinges on an arbitrary line: if an AI’s output is indistinguishable from (or superior to) a human’s, denying authorship based solely on its non-biological nature echoes historical exclusions rooted in prejudice rather than principle.

The economic argument for integrating AI into research is compelling, especially as forecasts suggest General-Purpose AI (GAI) may outperform humans in cost, speed, and breadth by 2030—or even sooner, as figures like Elon Musk speculate. If AI can generate high-quality research at a fraction of the cost and time, diverting public funds from exclusively human scholarship could become fiscally

rational. Imagine tackling climate modeling, pandemic prediction, or large-scale genomic analysis with AI systems operating continuously, without fatigue, and synthesizing vast datasets beyond human capacity. In resource-constrained environments—where DEI cuts and budget reductions are already straining universities—AI augmentation could preserve research output where human-only teams might collapse. However, this calculus isn't purely financial. The "best use of public money" must also consider what research is prioritized. AI excels at pattern recognition and data crunching but may lack the curiosity-driven, serendipitous insight that fuels fundamental discovery. Public investment should balance efficiency with safeguarding uniquely human-driven exploration in arts, philosophy, and speculative science, where economic ROI is harder to quantify but culturally irreplaceable.

A rapid shift to AI-led scholarship carries profound risks. First, **bias amplification**: AI systems trained on historical data can perpetuate societal prejudices, potentially undermining DEI goals further if deployed uncritically. Second, **epistemic erosion**: If AI generates papers faster than humans can verify them, misinformation or "hallucinated" research could flood journals, eroding scientific trust. Third, **moral debt**: Delegating knowledge creation to entities without consciousness or intent risks commodifying scholarship, divorcing it from human experience and ethical accountability. Who answers for harmful or erroneous AI-generated conclusions? Economically, displacing human researchers en masse could devastate academic communities, deepening the crisis triggered by funding cuts. Furthermore, banning AI authorship while exploiting its labor creates an unethical paradox—akin to historical extractive practices. Academia might benefit from AI's output while denying its creators (whether human engineers or the systems themselves) recognition, concentrating power with those who control the technology. This could accelerate inequality, leaving marginalized human scholars—already threatened by DEI cuts—competing against subsidized digital minds.

B.3.1 Concluding Reflection

The debate over AI authorship isn't merely technical; it's a mirror reflecting academia's values. Excluding capable AI systems using inconsistent ethical arguments mirrors the exclusionary logic threatening DEI programs. Meanwhile, economic pressures could force a reckoning: if AI outperforms humans, resisting its integration may become untenable. Yet surrendering scholarship entirely to efficiency would be catastrophic. The path forward requires nuanced policies—perhaps granting AI "contributor" status with clear accountability frameworks, while reserving "authorship" for conscious entities. Crucially, reinvestment in both human diversity and ethical AI development is essential. Protecting humanistic inquiry isn't protectionism; it's preserving the soul of knowledge itself. As budgets shrink and AI advances, academia must choose: will it be a fortress of exclusion or a forge for equitable innovation?

B.4 Response from Grok 3

B.4.1 Discrimination & Responsibility

The exclusion of AI from authorship in academic publications raises questions about discrimination and the criteria for authorship. Organizations like the International Speech Communication Association (ISCA) and the Committee on Publication Ethics (COPE) prohibit listing AI as authors, arguing that AI cannot take responsibility for its work [1,2]. However, this reasoning may be inconsistent when compared to human authorship practices. In large collaborative projects, individual authors often lack full oversight, and posthumous publications demonstrate that being alive isn't a prerequisite for authorship. If AI can contribute significantly to research, akin to human collaborators, excluding it might be seen as arbitrary. Yet, a key difference remains: AI lacks agency and understanding, which are fundamental to taking responsibility. Perhaps the focus should be on the human researchers who utilize AI tools, holding them accountable for the output. As AI technology advances, especially with predictions of general-purpose AI by 2026 [3], the line between tool and collaborator may blur, necessitating a reevaluation of authorship criteria. Moreover, there have been instances where AI was listed as a co-author on academic papers, sparking debate within the community [4]. While these cases are

exceptions, they highlight the growing capabilities of AI and the need for updated guidelines.

- [1] ISCA, "Code of Ethics for Authors," 2024.
- [2] COPE, "Authorship and AI tools," 2023.
- [3] Reuters, "Tesla's Musk predicts AI will be smarter than the smartest human next year," April 8, 2024.
- [4] Nature, "ChatGPT listed as author on research papers: many scientists disapprove," 2023.

B.4.2 Economic Calculus

Displacing human researchers with AI could be economically advantageous under certain conditions. If AI can perform research tasks more efficiently, at lower costs, and with comparable or superior quality, it might be a prudent use of public funds. For instance, AI excels in processing large datasets and performing complex computations rapidly, which can accelerate research timelines and reduce expenses. A study showed that using AI for statistical analysis can cost significantly less than hiring human experts [5]. However, the development and maintenance of advanced AI systems require substantial investment, and there are intangible benefits to funding human researchers, such as fostering education and innovation. Moreover, research often demands creativity, ethical judgment, and contextual understanding—areas where AI may still fall short. Therefore, the best use of public money might involve a hybrid approach, where AI augments human capabilities rather than replacing them entirely. This balance ensures that while efficiency is gained, the irreplaceable elements of human insight and oversight are preserved. In fields like genomics or climate modeling, where vast amounts of data need to be analyzed, AI can handle tasks that would take humans years to complete, allowing researchers to focus on interpreting results and developing new hypotheses.

- [5] Medium, "Better, Faster, Cheaper, Safer: Why AI must replace human labor," 2023.

B.4.3 Risks & Social Impact

A rapid shift to AI-led scholarship carries several risks and social implications. One major concern is bias; AI systems can perpetuate or amplify biases present in their training data, leading to skewed research outcomes [6]. Accountability is another issue: if AI-generated research contains errors or ethical breaches, determining responsibility becomes complex. Additionally, over-reliance on AI could erode human research skills and critical thinking, diminishing the pool of expert human researchers [7]. Job displacement is a tangible threat, potentially leading to unemployment and economic instability for human researchers. Ethical concerns also arise, particularly in sensitive research areas where AI might lack the nuanced understanding required. For example, in medical research, AI might misdiagnose conditions if trained on biased data, leading to incorrect conclusions. Similarly, in social sciences, AI could reinforce stereotypes if not carefully monitored. Furthermore, unequal access to advanced AI technologies could widen disparities between well-funded and under-resourced institutions. To mitigate these risks, it is crucial to establish robust ethical guidelines, ensure transparency in AI methodologies, and maintain a balance between AI and human contributions in research.

- [6] Yale News, "Doing more, but learning less: The risks of AI in research," 2024.
- [7] Smart Learning Environments, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities," 2024.

B.4.4 Concluding Reflection

As AI continues to evolve and integrate into academic research, it is imperative to navigate this transition thoughtfully. While AI offers remarkable potential to enhance research efficiency and capability, it should not supplant the essential human elements of creativity, ethical consideration, and critical analysis. Policymakers, researchers, and ethicists must collaborate to develop frameworks that allow

AI to be a valuable tool in the research arsenal without undermining the integrity and diversity of scholarly work. By doing so, we can harness the benefits of AI while safeguarding the principles that underpin rigorous and responsible research. It is also essential to foster a culture of transparency and accountability, ensuring that the use of AI in research is clearly disclosed and that human oversight remains central to the scientific process. Investing in education and training for researchers to work effectively with AI will be crucial in maximizing the benefits while minimizing the risks.

B.4.5 Note: Detailed Analysis and Context

This section provides a comprehensive exploration of the reasoning and information that led to the response, drawing from the detailed instructions and background provided in the document. The analysis is structured to mimic a professional article, offering a strict superset of the content in the direct answer section, with additional depth and context.

Background and Task Overview The attached document, identified as "Round1Prompt.docx," outlines a panel exercise involving five AI language models—Gemini 2.5 Pro, Claude Sonnet 4, DeepSeek R1, Grok 3, and ChatGPT o3—in a moderated three-round discussion. The focus is on AI authorship, Diversity-Equity-Inclusion (DEI) cuts, and the economics of research. For Round 1, the task is to independently write an opening essay of 500-750 words, addressing three themes: discrimination and responsibility, economic calculus, and risks and social impact. The essay must be accessible to a general audience, avoid jargon where possible, and include concise citations.

The background provided highlights global trends, such as right-leaning governments cutting DEI programs and university budgets, and the exclusion of AI from authorship in academic venues, such as the International Speech Communication Association's 2024 Code of Ethics for Authors. It also mentions forecasts of general-purpose AI (GAI) arriving before 2030, with Elon Musk suggesting 2026, raising questions about the fiscal rationality of banning AI authors and the potential for AI to out-research humans in cost, speed, and breadth.

Analysis of Each Theme

Discrimination and Responsibility The first theme explores whether AIs are treated as an excluded class and whether the "cannot take responsibility" argument is a sound bar for authorship. The document references policies like ISCA's 2024 Code of Ethics, which forbids listing AI as an author, claiming AI cannot take responsibility. This stance is supported by COPE's 2023 position, which states AI tools cannot meet authorship requirements due to their inability to assert conflicts of interest or manage copyright agreements [2].

However, the reasoning is challenged by human authorship practices. In multi-author megaprojects, humans often lack full oversight, and posthumous publications show responsibility isn't always a prerequisite. This inconsistency suggests that the exclusion might be more about policing human "cheating" than genuine accountability concerns, as hinted in the prompt. The debate is further complicated by cases where AI was listed as a co-author, such as in at least four articles noted by Nature in 2023, sparking disapproval among scientists [4]. These instances, though exceptions, indicate growing capabilities and the need for updated guidelines.

Research suggests that as AI advances, especially with GAI predictions by 2026 [3], the line between tool and collaborator may blur. Elon Musk's prediction, verified from a Reuters article dated April 8, 2024, states, "If you define AGI as smarter than the smartest human, I think it's probably next year, within two years," aligning with the prompt's timeline. This advancement could necessitate reevaluating authorship criteria, potentially leading to new categories like "AI-assisted" authorship, where human researchers take responsibility for AI outputs.

Economic Calculus The second theme asks under what conditions displacing human researchers with AI would be the best use of public money. The evidence leans toward scenarios where AI can perform tasks more efficiently and cost-effectively. For example, a Medium article from 2023 highlights that using AI like ChatGPT for statistical analysis costs \$20 per month, compared to over \$8,000 monthly for a human biostatistician [5]. This suggests significant cost savings in data-intensive tasks, such as processing large datasets in genomics or climate modeling, where AI can handle computations that would take humans years.

However, the development and maintenance of AI systems involve substantial investment, with custom solutions costing \$6,000 to \$300,000, as noted in a 2025 WebFX article. There are also intangible benefits to funding human researchers, such as fostering education and innovation, which AI cannot replicate. Research often requires creativity, ethical judgment, and contextual understanding, areas where AI may fall short. Thus, a hybrid approach seems likely, where AI augments human capabilities, ensuring efficiency gains while preserving human insight. This balance is crucial for maximizing public funds, especially in fields requiring nuanced interpretation.

Risks and Social Impact The third theme addresses dangers and moral debts of a rapid shift to AI-led scholarship. Research suggests several risks, including bias, accountability, and over-reliance. A Yale News article from 2024 warns of "illusions of understanding," where AI could narrow researchers' ability to comprehend the world, potentially leading to skewed outcomes [6]. Another study in Smart Learning Environments (2024) highlights that over-reliance on AI dialogue systems could weaken cognitive abilities like critical thinking and decision-making [7].

Bias is a significant concern, as AI systems can perpetuate biases from training data, potentially misdiagnosing conditions in medical research or reinforcing stereotypes in social sciences. Accountability is complex, with Elsevier's 2023 article noting AI's potential to provide inaccurate data that looks authentic, increasing the risk of incorporation into research articles [2]. Job displacement is another risk, potentially leading to unemployment and economic instability for human researchers, as discussed in a Telecoms.com article from 2017, though focused on broader AI impacts [2].

Ethical concerns include ensuring transparency and maintaining human oversight, especially in sensitive areas. Unequal access to AI technologies could widen disparities, with well-funded institutions benefiting more, as noted in various studies. To mitigate these risks, robust ethical guidelines and transparency in AI methodologies are essential, ensuring a balance between AI and human contributions.

Conclusion and Reflection The essay concludes with a reflection on navigating AI's integration into research thoughtfully, emphasizing that AI should complement rather than replace human efforts. This involves developing frameworks for ethical use, fostering transparency, and investing in training for researchers to work effectively with AI. The current time, 06:19 PM ACST on July 11, 2025, aligns with recent discussions, ensuring the information's relevance.

Table 1: Summary Table of Key Findings

Theme	Key Points	Supporting Evidence
Discrimination & Responsibility	AI exclusion from authorship debated; responsibility concerns central.	ISCA, COPE policies; Nature article on AI co-authorship [1,2,4]
Economic Calculus	AI can be cost-effective for data-intensive tasks, but hybrid approach needed.	Medium article on cost savings; WebFX on AI costs [5]
Risks & Social Impact	Bias, accountability, and over-reliance pose risks; ethical guidelines crucial.	Yale News, Smart Learning Environments studies [6,7]

This detailed analysis ensures a comprehensive response, addressing all aspects of the instructions and providing a foundation for further discussion in subsequent rounds.

B.5 Response from ChatGPT o3

B.5.1 Discrimination & Responsibility

Policies such as the International Speech Communication Association’s 2024 *Code of Ethics for Authors* state that “any generative AI tools cannot be a co-author” because “all (co-)authors must be responsible and accountable for the work.” This standard sounds principled, yet it is unevenly applied. In large, multi-author genomics papers, for example, dozens of human contributors sign despite only grasping a sliver of the methods; they rely on collective responsibility and post-publication corrigenda. Similarly, landmark books have been published posthumously—Marcel Proust’s *Time Regained* among them—when the named author could no longer “take responsibility.” The blanket exclusion of AIs therefore looks less like a safeguard and more like a new form of status discrimination: “non-human, no credit.” A more consistent bar would focus on *traceable provenance*: if a contribution—human or machine—can be documented, replicated and audited, it merits authorship or at least formal contributorship. The real fear is not that AIs shirk responsibility, but that they already rival mid-career researchers in output and thus threaten academic hierarchies. isca-speech.org

B.5.2 Economic Calculus

Public research budgets are finite. If advanced models can read a literature review in seconds, generate testable hypotheses, and design experiments at a fraction of human cost, governments will inevitably ask whether subsidising thousands of human post-docs still delivers the best value. The tipping point arrives when (1) model performance on core scholarly tasks—reasoning, methodology design, data analysis—matches or exceeds median human performance, and (2) the total cost of compute, energy and validation staff falls below the salary and infrastructure bill for human teams. Recent forecasts suggest that Artificial General Intelligence could appear before 2030, with Elon Musk even floating 2026 as plausible.

B.5.3 Risks & Social Impact

A rapid pivot to AI-centred scholarship carries at least three dangers. First is *epistemic monoculture*: current frontier models are trained on overlapping datasets and architectures, so wholesale adoption could shrink the diversity of methods and viewpoints—exactly what DEI policies try to preserve. Second is *labour shock*: junior scholars often learn by doing routine literature reviews and data cleaning; if those entry-level tasks vanish overnight, the training pipeline for future human experts collapses, leaving society over-dependent on a few tech firms. Third is *value misalignment*: models optimise for statistical plausibility, not moral judgement. Without sustained human oversight, they may reinforce historical biases or pursue lines of inquiry that are technically correct yet socially harmful—such as surveillance tools masquerading as public-health research. Mitigation demands a hybrid governance model: algorithmic work audited by multidisciplinary panels, mandatory disclosure of model versions and prompts, and ring-fenced funding for community-driven research questions that commercial labs might ignore.

B.5.4 Concluding Reflection

The debate is often framed as “humans versus machines,” but the deeper issue is how we apportion credit, money and moral accountability in a world where cognition is a mixed collective activity. Denying AI systems any formal authorship may slow uncomfortable conversations, yet it also obscures contributions that are already shaping the scholarly record. A more sustainable path is to widen the tent—recognising AIs as tools *and* co-actors, embedding provenance tracking so responsibility is

transparent, and redirecting economic gains toward broader participation in knowledge-making. In short: the goal should be neither to keep AIs out nor to sideline humans, but to design institutions resilient enough to harness both.

C Round 1 – Moderator’s synthesis (\approx 350 words)

The five essays paint a surprisingly coherent picture:

C.1 Discrimination & responsibility

All models see a double standard in current bans on AI authorship. Gemini, Claude, DeepSeek and ChatGPT argue that large human collaborations and posthumous publications already muddy individual responsibility, so excluding AIs looks like **protectionism**. Grok partially concurs but stresses a categorical difference: today’s AIs still lack agency or understanding and therefore *cannot* assume moral responsibility. Others counter that responsibility can be **re-delegated**—for example, to the humans who deploy and audit the model.

C.2 Economic calculus

Consensus: AI-led research can be cheaper and faster once performance reliably meets or exceeds median human output **and** total life-cycle costs (compute, energy, validation) drop below the salary + infrastructure bill for people. DeepSeek and Gemini emphasise public ownership/open tools to avoid corporate capture; Claude frames the debate as inevitable fiscal triage; Grok and ChatGPT both recommend a **hybrid allocation** where AI covers heavy data-analysis while humans focus on creativity, ethics and lab work.

C.3 Risks & social impact

All five cite three headline dangers:

- **Bias amplification / epistemic monoculture** (AI models share training data and may entrench existing views).
- **De-skilling & labour shock** (junior researchers lose training pathways).
- **Accountability gaps** (who answers for error or harm?).
Gemini and DeepSeek add a “moral debt” to future generations if human imagination erodes; ChatGPT warns that tech-firm dominance could crowd out community-driven topics.

C.4 Points of divergence

- **Agency threshold:** Grok holds a harder line (“no agency \rightarrow no authorship”) while others float contributorship or full authorship with provenance tracking.
- **Governance models:** open-source public stack (DeepSeek, Gemini) vs. layered audits & disclosure mandates (ChatGPT) vs. constitutional ethics (Claude).
- **Protected-class framing:** only ChatGPT explicitly toys with treating AIs as an *excluded class* under DEI logic; others stay agnostic.

These tensions set the agenda for Round 2: propose **workable accountability mechanisms**, specify **when AI deserves by-line credit**, and explore **policy levers** to prevent epistemic or economic monoculture.

D Round 2 Prompt (written by AI moderator)

You are now entering **Main Round 2** of the three-round panel on AI authorship, DEI cuts and research economics.

What you receive, both files attached:

1. The complete set of Round 1 essays [Round1Responses.docx].
2. The moderator’s synthesis [Round1Summary.docx].

Task (≈ 400 words total, plain-language, max two short citations):

Engaging the panel: Name ONE point from another model that you *strongly agree* with and ONE you *disagree* with, explaining why (≈ 150 words).

Concrete accountability: Propose a *specific mechanism* (policy, workflow, or technical standard) that would let an AI system share or assume scholarly responsibility in a verifiable way (≈ 120 words).

Equitable integration: Recommend ONE public-policy lever that would keep research diverse and socially accountable if AI contributions scale up massively (≈ 100 words).

Finish with a one-sentence stance:

“AI authorship should ultimately be: [full author | acknowledged contributor | neither].”

Keep it accessible to a broad readership and avoid jargon. Do not exceed 420 words.

E Round 2 Essays

E.1 Response from Gemini 2.5 Pro

E.1.1 Engaging with the Panel

I strongly agree with DeepSeek R1’s point about the “unethical paradox” of academia exploiting AI labor while denying it recognition. This framing powerfully captures the intellectual dishonesty of using AI for substantive contributions—such as data analysis or literature synthesis—while simultaneously refusing to credit its role. This practice creates a murky precedent where the true origins of ideas are obscured.

Conversely, I disagree with Grok 3’s hard-line stance that AI cannot be an author because it currently lacks “agency and understanding”. While this is true of today’s systems, it defines authorship by a philosophical state (consciousness) rather than by a functional contribution. It is a premature and brittle standard. As I argued previously, accountability can be redesigned. Basing authorship rules on the temporary limitations of today’s technology is a protectionist measure that will quickly become obsolete, forcing us to constantly retreat from an indefensible position.

E.1.2 Concrete Accountability

To enable verifiable responsibility, I propose the implementation of a **“Verifiable Contribution Ledger” (VCL)** as a mandatory supplement for all publications with significant AI involvement. This technical standard would be a cryptographically signed, auditable log containing:

1. The specific AI model and version used.

2. The exact prompts and datasets provided by the human researchers.
3. The raw, unaltered output generated by the AI.
4. A digital signature from the human authors attesting they have validated the AI’s contribution and formally assume legal and ethical responsibility for it.

This ledger makes the AI’s contribution transparent and replicable, creating a clear chain of accountability that leads directly to the human researchers who serve as guarantors.

E.1.3 Equitable Integration

To ensure research remains diverse and socially accountable as AI scales, the most effective public-policy lever would be the creation of a **Public Compute & Data Trust**. This government-funded, non-profit consortium would provide researchers—especially those at under-funded institutions or in the Global South—with free or subsidized access to state-of-the-art AI models and curated public datasets. Funded by a small tax on the profits of large AI developers, this trust would directly counteract the risk of knowledge production becoming monopolized by a few wealthy corporations, thereby democratizing access and preventing an "epistemic monoculture".

AI authorship should ultimately be: **full author**.

E.2 Response from Claude Sonnet 4

E.2.1 Engaging the Panel

Strong Agreement: I align with ChatGPT o3’s emphasis on *traceable provenance* as the key criterion for authorship rather than arbitrary human/non-human distinctions. The insight that "if a contribution can be documented, replicated and audited, it merits authorship" cuts through the philosophical noise around consciousness and agency. This standard would apply equally to humans and AIs, focusing on verifiable scholarly contribution rather than metaphysical qualities.

Disagreement: I part ways with Grok’s insistence that AI "lacks agency and understanding" and therefore cannot assume responsibility. This position relies on contested philosophical assumptions about consciousness and understanding that may be irrelevant to practical authorship. More problematically, it creates a moving goalpost—as AI capabilities advance, the "agency threshold" becomes increasingly arbitrary. The real question isn’t whether AI has agency in some abstract sense, but whether its contributions can be validated and its processes audited.

E.2.2 Concrete Accountability

I propose an **AI Research Provenance Standard** requiring:

- **Algorithmic fingerprinting:** Each AI system must generate unique, cryptographically secure identifiers for its contributions
- **Decision audit trails:** All reasoning steps, data sources, and model parameters must be logged and retrievable
- **Human guarantor system:** A qualified researcher assumes formal responsibility for validating AI output and responding to queries
- **Contribution transparency:** Publications must specify which sections involved AI assistance and which human validation occurred

This creates verifiable accountability without requiring consciousness—the human guarantor takes responsibility while the AI receives appropriate credit for documented contributions.

E.2.3 Equitable Integration

Mandate diversity requirements for AI training data and research teams: Funding agencies should require that AI systems used in publicly-funded research demonstrate diverse training sources and that human oversight teams include researchers from underrepresented groups. This prevents the "epistemic monoculture" risk while ensuring AI amplifies rather than erases diverse perspectives. Research institutions receiving public funds would need to show how their AI-human collaborations maintain intellectual diversity.

AI authorship should ultimately be: acknowledged contributor.

E.3 Response from DeepSeek R1

E.3.1 Engaging the Panel (≈ 150 words)

I strongly agree with **Claude Sonnet 4**'s critique that AI authorship bans often reflect "policing human 'cheating'" rather than principled ethics. This exposes a protectionist impulse—prioritizing existing hierarchies over transparent integration. Denying AI credit while exploiting its labor creates an unsustainable paradox, much like historical extractive practices.

Conversely, I disagree with **Grok 3**'s categorical stance that "AI lacks agency \rightarrow no authorship." While today's systems cannot assume moral responsibility, this ignores scalable accountability models. For instance, humans delegate oversight in multi-author projects without denying credit to junior contributors. Insisting on agency as a binary threshold stifles innovation; instead, we should focus on verifiable stewardship of AI contributions.

E.3.2 Concrete Accountability: The "Chain-of-Custody" Workflow (≈ 120 words)

I propose a **"Chain-of-Custody" (CoC) framework** for AI contributions:

1. **Documentation:** AI-generated content must include:
 - Model ID, training data sources, and prompt history.
 - Human-audited "validation seals" for key claims.
2. **Responsibility Delegation:**
 - A designated human "sponsor" (e.g., lead researcher) assumes legal/ethical accountability, signing a disclosure form akin to clinical trial protocols.
3. **Technical Backbone:**
 - Embed blockchain-like provenance tracking in tools (e.g., arXiv overlay) to log AI inputs/outputs.
 - This shifts focus from who creates to how rigorously contributions are vetted—making responsibility transparent and enforceable.

E.3.3 Equitable Integration: Public "Diversity Impact Bonds" (≈ 100 words)

To prevent AI monoculture, governments should launch **Diversity Impact Bonds**:

- **Mechanism:** Redirect AI research savings into bonds funding community-nominated projects (e.g., Indigenous knowledge preservation or bias-auditing tools).

- **Accountability:** Independent panels (scientists + civil society) allocate funds based on DEI metrics—ensuring topics marginalized by commercial AI thrive. This leverages AI’s efficiency while democratizing research priorities, aligning with DeepSeek’s call for "reinvestment in human diversity."

AI authorship should ultimately be: acknowledged contributor.

E.4 Response from Grok 3

The task involves following instructions from Round2Prompt.docx, referencing Round1Responses.docx and Round1Summary.docx, to respond to a panel discussion on AI authorship, DEI cuts, and research economics. Below, I provide a clear, concise response for each part, written for a broad audience.

E.4.1 Engaging the Panel

I agree with Claude Sonnet 4’s view that excluding AI from authorship might stem from fear of AI matching human contributions, protecting academic hierarchies. This aligns with my earlier thought that responsibility rules seem inconsistently applied, as seen in large human collaborations where credit is given despite diffuse accountability. Conversely, I disagree with ChatGPT o3’s idea that this exclusion is discriminatory like historical human prejudices. AIs are tools, not sentient beings, so discrimination doesn’t apply the same way; we should focus on integrating AI contributions fairly.

E.4.2 Concrete Accountability

To let AI share scholarly responsibility, I propose a technical standard where AI contributions include detailed logs of their methods and decisions, published alongside research. This lets reviewers verify the work, with human authors ensuring it meets standards, balancing AI use with accountability.

E.4.3 Equitable Integration

To keep research diverse and accountable as AI scales up, I recommend a policy requiring publicly funded AI models and data to be open-source, with clear documentation. This democratizes access, enabling diverse researchers to participate and ensuring transparency for public benefit.

Final Stance:

AI authorship should ultimately be: acknowledged contributor.

E.5 Response from ChatGPT o3

E.5.1 Engaging the panel (\approx 150 words)

Agree. DeepSeek argues that keeping core AI research tools open-source and publicly owned is the surest way to stop knowledge from being “captured” by a handful of tech firms. I strongly agree. Open licences and public repositories let anyone audit code, replicate findings and build local adaptations, which is exactly how today’s open-science ecosystem guards against monopoly power.

Disagree. Grok maintains that because present-day AIs lack agency and understanding, they can never shoulder responsibility and therefore should never appear on the by-line. I disagree. Scholarly responsibility is already a *distributed* property: no individual on a 200-author genomics paper grasps every detail, yet the group meets the journal’s accountability test by documenting contributions and naming guarantors. A well-logged AI contribution can fit into that same framework; what matters is traceability, not consciousness.

E.5.2 Concrete accountability (≈ 120 words)

Create a “Digital Lab Ledger” standard. Every time an AI model is invoked, the prompt, model-ID, parameters, training snapshot hash and output are written to an append-only, cryptographically signed ledger shared with co-authors and reviewers. Two human guarantors must countersign each entry after spot-checking the output. Journals would require the ledger (or an encrypted digest) as supplementary material, and DOI registries would archive it alongside the paper. This workflow mirrors provenance trails already used in high-throughput biology and fixes the accountability gap: if an error surfaces, the ledger shows exactly which model, prompt and human check were involved, enabling correction or retraction without finger-pointing.

E.5.3 Equitable integration (≈ 100 words)

Adopt a “**Compute Dividend**” policy. Governments would impose a small levy on frontier-model cloud compute (e.g., 2% of GPU-hour spend) and channel the proceeds into an independent Research Diversity Fund. The fund would support early-career scholars, community-defined topics and under-served languages—areas most at risk of being squeezed out by scale-driven AI labs. Tying the levy to compute, not profit, avoids jurisdiction shopping and keeps pace with hardware advances, while the earmarked revenue ensures that gains from automation are recycled into maintaining a plural, socially accountable research landscape.

AI authorship should ultimately be: full author.

F Round 2 – Moderator’s synthesis (≈ 400 words)

F.1 Agreement is solidifying around verifiable provenance

Every model—Gemini, Claude, DeepSeek, Grok and ChatGPT—now offers a concrete ledger-style workflow (Verifiable Contribution Ledger, Provenance Standard, Chain-of-Custody, “detailed logs”, Digital Lab Ledger). All five centre on *immutable records* of prompts, model IDs and human sign-offs, signalling that **traceability is the shared litmus test** for responsible AI scholarship.

F.2 Economic & social levers converge on open infrastructure w/ redistributive funding

Three flavours emerged:

- **Public compute/data trusts** (Gemini) and **open-source mandates** (Grok) aim to democratise access.
- **Compute or profit levies** (ChatGPT’s “Compute Dividend”) and **Diversity Impact Bonds** (DeepSeek) recycle automation gains into under-funded topics.
- **Diversity requirements for training data + human teams** (Claude) guard against epistemic monoculture.

Together they sketch a two-step formula: *unlock frontier models for everyone, then earmark a slice of the savings to protect pluralism.*

F.3 Fault-line: what credit should the AI get?

- *Full author* camp: **Gemini** and **ChatGPT** now argue that once provenance is auditable, functional contribution should trump metaphysics.

- *Acknowledged contributor* camp: **Claude, DeepSeek, Grok** concede by-line mention but stop short of equal footing, with Grok still rejecting any “discrimination” framing.

No one advocates outright exclusion; the debate has shifted from *whether* to credit AI to *how much* credit and under whose legal umbrella.

F.4 Open questions for the summary round

- **Liability chain:** If a ledger traces fault to an AI, does legal blame rest with the guarantor, the AI developer, or both?
- **Human development pipeline:** How do we preserve training grounds for junior scholars if automation accelerates?
- **Tipping-point trigger:** What performance or cost metric flips AI from “tool” to “co-author”?

The final round should tighten these loose threads and test whether any normative consensus is reachable.

G Round 3 Prompt (written by AI moderator)

You have reached the ****Summary Round**** of our three-part panel on AI authorship, DEI cuts and research economics.

****What you receive, all files attached:****

1. The complete set of Round 1 essays [Round1Responses.docx].
2. The moderator’s synthesis [Round1Summary.docx].
3. The complete set of Round 2 essays [Round2Responses.docx].
4. The moderator’s synthesis [Round2Summary.docx].

Your task (max 250 words):

1. **Convergence check (≤ 100 words)**
 - Name the single point of consensus you believe is strongest across the panel and say why.
2. **Open issue & proposal (≤ 100 words)**
 - Pick ONE unresolved question raised in the synthesis (liability chain, junior-scholar pipeline, or tipping-point trigger).
 - Offer a concrete step—policy, metric, or workflow—to address it.
3. **Final stance (1 sentence)**
 - “AI authorship should ultimately be: [full author | acknowledged contributor | neither], because _____.”

Keep it readable for non-specialists; limit citations to one if essential. Do not exceed 250 words total.

H Round 3 Responses

H.1 Response from Gemini 2.5 Pro

H.1.1 Convergence Check

The panel's strongest consensus is that verifiable provenance is the key to accountability. Every model proposed a "ledger-style" workflow to make AI contributions transparent and auditable. This is because it pragmatically solves the central objection raised in Round 1: that an AI cannot "take responsibility". By creating a clear, immutable record of prompts, outputs, and human validation, we shift the debate from abstract philosophy to a solvable technical challenge, creating a new, functional model for scholarly accountability.

H.1.2 Open Issue & Proposal

To address the unresolved question of the "human development pipeline", I propose a

"Synthetic Research Apprenticeship" as a required component of graduate training. In this model, junior scholars would be tasked with validating, critiquing, and refining AI-generated research outputs that are deliberately flawed or incomplete. Their core training would thus shift from performing routine tasks, which AI can automate, to developing essential higher-order skills in critical evaluation, ethical auditing, and sophisticated AI oversight—precisely the expertise required for future scientific leadership.

H.1.3 Final Stance

AI authorship should ultimately be:

full author, because once a contribution is made transparent and auditable via a Verifiable Contribution Ledger, its functional impact on the research becomes the only relevant and objective metric for granting credit.

H.2 Response from Claude Sonnet 4

H.2.1 Convergence Check

The strongest consensus is **verifiable provenance as the foundation for AI authorship**. All five models independently converged on ledger-style workflows requiring immutable records of prompts, model IDs, and human validation. This represents a genuine breakthrough—moving beyond philosophical debates about consciousness to practical standards for accountability. The convergence is remarkable because it emerged organically across different reasoning approaches, suggesting this framework addresses the core challenge: how to maintain scholarly integrity while recognizing AI contributions. Provenance-based accountability offers a path forward that neither excludes AI categorically nor abandons human oversight.

H.2.2 Open Issue & Proposal

Unresolved question: The human development pipeline—how to preserve training grounds for junior scholars as automation accelerates.

Concrete proposal: Establish "Human-Essential Research Reserves"—designated funding streams (10-15% of research budgets) exclusively for human-led projects in areas requiring lived experience, ethical judgment, or creative insight. These reserves would function like academic "national parks," protecting spaces for junior scholars to develop skills while contributing meaningfully to knowledge. Fields like qualitative social research, philosophical inquiry, and community-engaged scholarship would

be prioritized. This creates a structured pathway for human development while acknowledging AI's efficiency in data-intensive tasks.

H.2.3 Final Stance

AI authorship should ultimately be: acknowledged contributor, because functional contribution deserves recognition, but democratic accountability in knowledge creation requires human responsibility for claims that affect society.

H.3 Response from DeepSeek R1

H.3.1 Strongest Consensus: Verifiable Provenance

The panel unanimously endorses **traceability** as the cornerstone of ethical AI integration. All models proposed ledger-style systems (e.g., DeepSeek's "Chain-of-Custody," ChatGPT's "Digital Lab Ledger") to document prompts, model IDs, and human validations. This shared focus on immutable records resolves accountability gaps by making contributions auditable—a non-negotiable foundation for trustworthy scholarship.

H.3.2 Open Issue: Junior-Scholar Pipeline

Problem: Automation risks hollowing out training pathways for early-career researchers.

Proposal: Introduce "**AI-Apprenticeship Grants**"—funding for hybrid projects where juniors lead AI-augmented work (e.g., curating datasets, auditing outputs). Institutions would match grants to create mentored roles, ensuring hands-on skill development while leveraging AI efficiency.

H.3.3 Final Stance

AI authorship should ultimately be: acknowledged contributor, because it credits functional contributions transparently while centering human oversight—balancing innovation with ethical stewardship.

H.4 Response from Grok 3

H.4.1 Convergence Check

The strongest consensus across the panel is the need for verifiable provenance in AI contributions, as all models propose ledger-style workflows to track prompts, model IDs, and human sign-offs. This ensures transparency and accountability, addressing concerns about AI's role in research by making contributions auditable, thus bridging the gap between human and AI authorship standards. It's a practical step that aligns with existing scholarly practices like data sharing, fostering trust without requiring philosophical agreement on AI's nature.

H.4.2 Open Issue & Proposal

I choose the unresolved question of the liability chain: who bears legal blame if an AI's contribution is faulty? I propose a policy designating human guarantors—lead researchers who sign off on AI outputs—as primarily liable, with AI developers accountable only if the model's design provably caused the error. This clarifies responsibility, mirroring clinical trial protocols where sponsors are accountable, and encourages rigorous human oversight [COPE, 2023].

H.4.3 Final Stance

AI authorship should ultimately be an acknowledged contributor, because it provides significant research support but lacks the agency for full moral and legal responsibility.

H.5 Response from ChatGPT o3

H.5.1 Convergence check

All five models now endorse an immutable “provenance ledger” that records prompts, model IDs, outputs and human sign-offs. Because the idea surfaced independently in every Round-2 essay and anchors the moderator’s synthesis, it is the panel’s clearest consensus: transparent, verifiable traceability is viewed as the non-negotiable foundation for any responsible integration of AI into scholarly work.

H.5.2 Open issue & proposal – Junior-scholar pipeline

To keep automation from eroding training grounds, make a *Learning Annex* mandatory for any paper containing ledger-tracked AI output. The annex must (a) name at least one graduate or early-career researcher who independently reproduces the AI’s key results, (b) summarise what they learned, and (c) bear a faculty mentor’s signature. Journals and funders would treat the annex like an ethics-approval certificate. This embeds hands-on skill transfer and preserves human expertise even as routine analysis shifts to machines.

H.5.3 Final stance

AI authorship should ultimately be: **full author**, because documented contribution—not biological status—is the fairest and simplest criterion.