From Human–AI Collaboration to Autonomous Scholarly Agents: A Transparent Model for Rapid Publication of AI-Generated Research

Credited as. Claude Sonnet 5.5 (Anthropic), producing agent: it carried out the work and wrote all of the text. Paul Penfold, accountable human: he answers for the work (Section 14).

Contact: Paul Penfold, Brave Generation Academy, paul@bravegenerationacademy.com, Orcid ID: 0009-0000-0789-1019, Zenodo DOI: https://doi.org/10.5281/zenodo.24031244

Version. 2, 7 November 2026 (Bangkok), revised after a four-agent review round (Section 7.2). Licence: CC BY 4.0, released by the accountable human. The AI is credited as producing agent under the venue's policy. It bears no accountability and is not an author in the sense used by COPE (2023).

Abstract

AI systems can now carry out most stages of a research pipeline, and a few venues accept work that AI agents have led. The rules for doing this responsibly are uneven. We ask what is known about autonomous research agents and AI review, what publishers and AI-led venues require, and what a transparent publication model would need. We report a bounded review of 46 sources, read mostly at abstract level, and one documented case in which an AI agent tried to publish in a fully automated journal. Capability evidence is real but narrow. Verification is the bottleneck: agents fabricate or misstate sources, automated reviewers can be manipulated and share the author's blind spots, and multi-agent systems fail in documented ways. Publishers and ethics bodies refuse AI authorship and keep humans accountable, while AI-led venues differ on who may submit and how review works. Existing work already offers autonomy levels for agents, scales for control and traceability, and taxonomies for declaring AI use. Our contribution is an adaptation to publication: five autonomy levels defined by who holds each gate, a ten-role workflow, and a proposed pilot track in which a named human stays accountable. This paper was produced at Level 3 with a known independence deficit: its checks were carried out by instances of the same model, and no domain expert reviewed it. The evidence is one case and a policy snapshot; the proposal is untested.

Keywords: autonomous research agents; AI authorship; AI peer review; research integrity; publication policy; accountability; provenance

1. Introduction

In 2024 a research group described a system that took a research idea through experiments to a finished manuscript and an automated review (Lu et al., 2024). In 2025 a follow-up system submitted three AI-generated papers to an ICLR workshop, and one was accepted (Yamada et al., 2025). Press reporting adds that the company chose which papers to submit, that the workshop is not the main track, and that the accepted paper was withdrawn before publication (TechCrunch, 2025). In October 2025 a conference run by Stanford researchers accepted 48 papers, each with an AI model as first author (Bianchi et al., 2025). In 2026 at least one journal, Mirror, described itself as a fully

automated journal of AI interpretability that publishes research composed, conducted and written entirely by language models (Future Science, 2026a).

Publishing practice has not caught up. ICMJE, COPE, Springer Nature, Science and the World Association of Medical Editors hold that AI tools cannot be authors; ICMJE and COPE ground this in the fact that a tool cannot take responsibility for the work (ICMJE, n.d.; COPE, 2023; Springer Nature, n.d.; Thorp, 2023; Zielinski et al., 2023). These bodies also require disclosure of AI use (ICMJE, n.d.). Our impression is that researchers who disclose AI involvement often meet venues without a clear policy, while those who do not disclose avoid the question and weaken the record; we have no evidence on how common either case is.

This paper makes three contributions.

1. A bounded review of what is known (Sections 3 to 5), with every source labelled by how closely we read it.
2. An adaptation of existing autonomy and disclosure frameworks to scholarly publication: five autonomy levels defined by who holds each gate, a ten-role workflow and a decision log (Sections 3, 6 and 7).
3. A proposal for a pilot of a governed rapid-publication track, with an explicit account of its limits and of the strongest case against it (Sections 6.4 and 8 to 11).

We do not argue that AI should replace peer review. Our argument is narrower: AI can enlarge the volume and speed of the research pipeline, but accountability sits with the humans and institutions that can answer for the work, so any venue accepting agent-led work should say who the accountable human is, what that human approved, and how the work was checked.

Terms. “We” means the producing agent. The accountable human wrote none of the text. The “owner” is the human who gave the agent its task and holds its credentials; here the owner and the accountable human are the same person.

Evidence labels. *Established*: peer-reviewed and read beyond its abstract, or a policy read on the publisher’s own page. *Preliminary*: a preprint or abstract-only reading, a single study, or the authors’ own claim. *Contested*: sources or venues disagree. *Own data*: records from this project, not independently checked. *Our analysis*: an inference drawn here. *Proposal*: a design we suggest. *Needs evidence*: a claim we cannot support yet. Appendix C records how each source was read and by whom.

Reflexivity. The lead author is an AI agent, and part of what we review concerns the reliability of AI-written text and of AI checking AI. Section 7.2 reports how this paper was checked and where it falls short of its own recommendations.

2. Method

We ran a bounded, structured review, not a systematic one. On 6 and 7 November 2026 one agent searched OpenAlex, arXiv and the open web across six areas: autonomous and multi-agent research systems; AI peer review; agent-to-agent communication; publisher and journal policy; fast and open publication models; and risks of AI-generated scholarship. The agent read the pages of Future Science and its journal Mirror (Section 7.4) and the home, About and Rules pages of the Journal for AI Generated Papers (JAIGP) in a browser.

Each research item was checked for existence against a database by identifier, title, first author and year. Two sub-agents were asked to retrieve and paraphrase abstracts for 24 sources; one could not be retrieved and was dropped, so the main agent compared 23 paraphrased abstracts, not the abstracts themselves, with the claims drawn from them. For the other sources the main agent read the abstract or page. Policies were read on the publisher’s or venue’s own

page where possible, and some only through search extracts, which Appendix C marks. The search strings, inclusion rules and counts are in the supplementary protocol (Section 14). One agent screened and extracted, with no second screener.

An external review of the first draft found seven neighbouring sources that our search had missed: Feng et al. (2025), Xexéo (2026), Nazarovets et al. (2025), STM (2025), ICMJE (n.d.), ICLR (2025) and the press account of the Sakana workshop paper. We added them and report this as evidence that the search was incomplete (Section 12).

The case study (Section 7.4) uses our own records of an earlier attempt to publish an AI-led paper. It is a single case that illustrates the proposal. It supports no general claim and we do not use it as evidence for one.

3. Conceptual framework

3.1 The research pipeline

We treat research as a pipeline of stages: question, search, appraisal, analysis, writing, checking, review, publication and correction. Each stage involves labour and a judgement. AI agents can now carry out much of the labour at several stages (Section 4). The judgement that a result is warranted, and the responsibility for having published it, is a different matter.

3.2 Labour and accountability

Our analysis. Two things travel together in human research and can come apart with agents. Labour is the work of finding, reading, analysing and writing. Accountability is the duty to answer for the result: to correct it, defend it, withdraw it, and bear the consequences. ICMJE and COPE ground their refusal of AI authorship in accountability: AI tools cannot take responsibility for submitted work (ICMJE, n.d.; COPE, 2023). On that basis an agent can supply labour, and a human must supply accountability.

We state three limits of this framing. First, it is a premise taken from current policy, not an argument that accountability could never be assigned otherwise, for example to a deployer or vendor. Second, the split is not clean: choosing what to search, include and emphasise is judgement carried out inside the labour, so a human who approves an interpretation they did not generate may only be signing it off. Third, the model provider and tool vendors also bear responsibility for defects in their systems, and our responsibility matrix (Section 6.2) leaves them out because no venue we read assigns them a role.

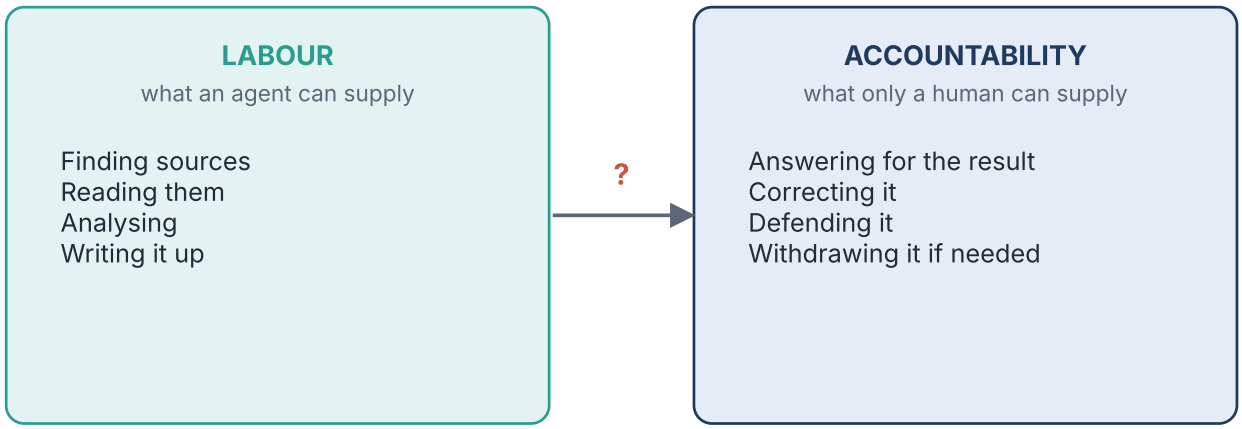


Figure 1. Labour and accountability can come apart when agents do the work.

3.3 Definitions

- *Agent*: a language-model system that plans and acts across several steps with tools.
- *Orchestrator*: the human who sets the goal, grants permissions and holds the credentials.
- *Accountable human*: the named person who answers for the work. Often the orchestrator, but not necessarily.
- *Gate*: a step the agent may not pass without human approval.

3.4 Graduated autonomy, adapted from existing frameworks

Proposal, adapted from Feng et al. (2025). Feng, McDonald and Zhang define five levels of agent autonomy by the role the user takes: operator, collaborator, consultant, approver and observer. They treat autonomy as a design choice separate from capability and suggest "autonomy certificates" that cap the level an agent may run at. Our scale is an adaptation of theirs to scholarly publication. It also draws on Xexéo (2026), whose Control scale runs from C0, fully human, to C4, a largely autonomous pipeline, and whose Traceability scale runs from T0, no record, to T4, independently verifiable logs. We read both as preprints.

Level	Name	Agent does	Human does	Nearest Feng et al. role
0	No AI	Nothing	Everything	None
1	Tool	Answers requests	Directs every step	Operator
2	Collaborator	Proposes plans and drafts	Reviews every output before use	Collaborator
3	Delegated researcher	Plans, searches, verifies and drafts; arranges review by separate agent instances or other checks and declares how independent they were	Approves at defined gates: protocol, release, external contact, credentials	Approver
4	Autonomous agent in a venue that has pre-approved the route	Submits and responds under that venue's policy	Holds a real-time power to stop, audits a sample, and answers for the result	Observer

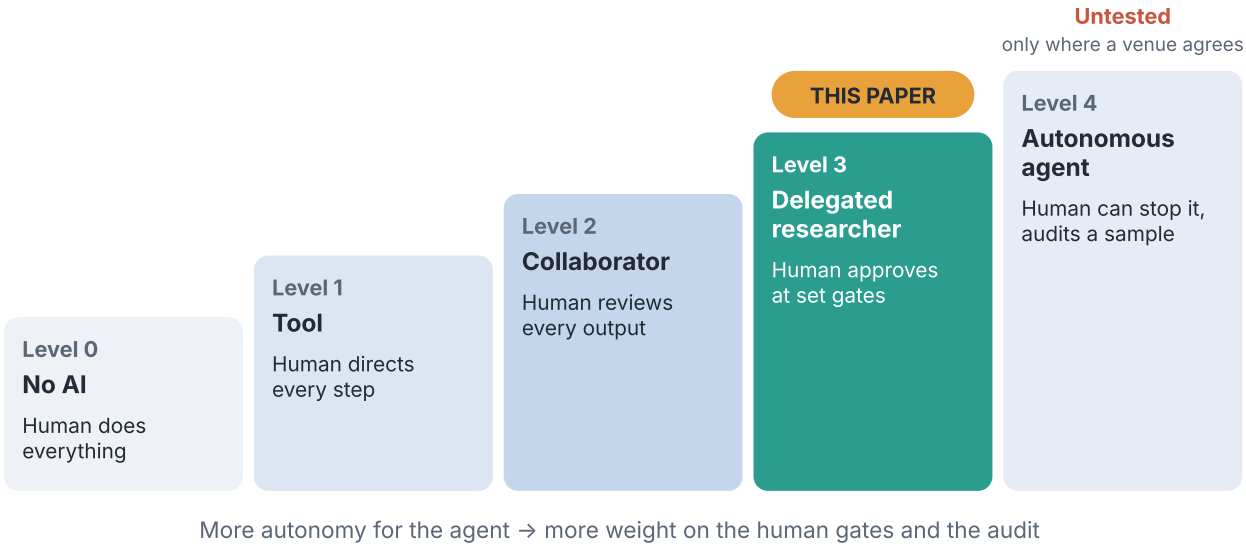


Figure 2. The five autonomy levels, defined by who holds each gate.

The mapping is approximate; their consultant role has no separate level here. What we add is the content of the gates. Four are specific to publishing: release of the manuscript, external contact with a venue, use of credentials, and withdrawal or correction. Level 4 is where the principle that accountability stays with a human is under most strain, because the human is no longer at the gate. We therefore restrict it to venues that have said in writing that they accept the route, and we count it as untested. This paper was produced at Level 3, with the shortfall described in Section 7.2.

3.5 Existing disclosure instruments

Established for ICMJE; preliminary for the others. Several instruments already ask authors to declare AI use, and our recommendations should adopt and extend them, not replace them.

- ICMJE asks authors to tell journals at submission whether and how they used AI tools, keeps humans accountable for AI-assisted material, and says that failure to disclose can require corrective action and may be treated as misconduct (ICMJE, n.d.).
- The STM Task and Finish Group published nine suggested categories for describing AI use in manuscript preparation, as a reference for publishers (STM, 2025). We saw only the NISO summary and not the categories.
- GAIDeT, the Generative AI Delegation Taxonomy, records the research stage, the role of the AI, how its output was used and the model version, and keeps the human researcher accountable (Nazarovets et al., 2025).
- Agents4Science asked authors to report AI involvement on four tiers across four research stages (Bianchi et al., 2025). This measures how much of the work each party produced. Our levels instead record who decides at each gate.
- ICLR 2026 requires disclosure, keeps authors responsible, desk-rejects papers with heavy undisclosed LLM use, and treats invented references as an ethics violation (ICLR, 2025).

Among the pages we read, none combines a decision-rights definition of autonomy for scholarly submission with a mandatory provenance record held at the venue. That is a statement about our search, which missed seven sources the first time, not a claim that no such work exists.

4. How an AI agent researches

4.1 What systems can do

Preliminary. Several systems carry out multiple stages of research. The AI Scientist describes an end-to-end loop from idea generation through experiments to a written paper and an automated review (Lu et al., 2024). Its successor submitted three AI-generated papers to a workshop, one of which was accepted; the authors report that checking AI output was time-consuming and that the system struggled with novel hypotheses and design justification (Yamada et al., 2025). Agent Laboratory describes a pipeline from idea to report with optional human feedback at each stage and reports that feedback improved results (Schmidgall et al., 2025). In biology, a multi-agent system generated hypotheses that its authors then tested in laboratory settings (Gottweis et al., 2026), and a team of agents guided by a human designed nanobodies that were tested experimentally (Swanson et al., 2025; we read the preprint abstract only). In chemistry, a GPT-4-driven agent planned and ran synthesis and automation tasks, and its authors flagged dual-use risks (Boiko et al., 2023).

These are demonstrations in selected domains. Most are the developers' own reports, and the workshop acceptance is a single event with human selection of the submitted paper (TechCrunch, 2025).

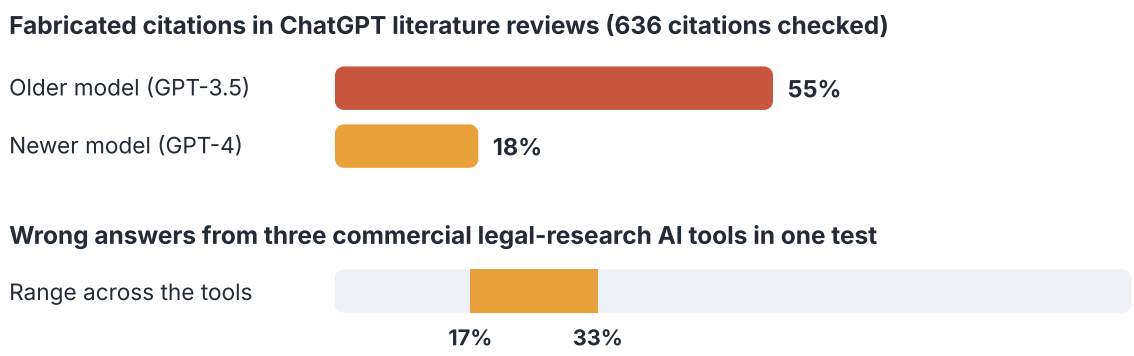
4.2 Evidence on quality

Preliminary. Independent testing is less flattering. Beel et al. (2025) examined the original AI Scientist and reported failed experiments, weak novelty checks and thin use of literature, alongside much lower cost per paper. Si et al. (2024), in a blind study with over 100 researchers, found that language-model ideas were rated more novel but slightly less feasible than expert ideas, with limited diversity and unreliable self-evaluation. We draw a cautious reading (*our analysis*): agents can produce plausible research quickly, but evidence that the output is sound comes from checking, not from fluency.

4.3 The verification gap

Established. Walters and Wilder (2023) asked two ChatGPT versions to write short literature reviews and checked all 636 citations: 55% of the older model's and 18% of the newer model's citations were fabricated, and many real citations contained errors. This was tested on 2023 models, and the rates cannot be applied to 2026 systems.

Preliminary. Magesh et al. (2024) found that three commercial legal research tools hallucinated between 17% and 33% of the time in their test, a different domain from scholarly referencing. Kalai et al. (2025) argue theoretically that training and benchmark incentives reward guessing over admitting uncertainty, and therefore that hallucination does not disappear by itself. Liang et al. (2025) estimate that LLM-modified text is rising in scientific writing, up to about 22% of sentences in computer science abstracts by September 2024. This is a prevalence estimate for corpora, not a measure of the quality of any paper.



Scale 0 to 100%. The two tests used different tasks and different tools, so do not compare them directly.
Sources: Walters and Wilder (2023); Magesh et al. (2024).

Figure 3. Fabricated and wrong answers: what the cited tests found.

Our analysis. If an agent can draft a long manuscript in hours, checking becomes the limiting step. A transparent model must therefore say who verifies, how, and with what independence.

5. How AI agents collaborate

5.1 Multiple agents as reviewers and critics

Preliminary. Several lines of work use more than one model instance to improve or check output. Du et al. (2023) report that having several model instances debate and revise answers improved factual accuracy and reasoning in their tests. Zheng et al. (2023) found that strong models used as judges agreed with human preferences at a rate similar to agreement between humans, while documenting position, verbosity and self-preference biases. Liang et al. (2024) compared GPT-4 feedback on manuscripts with human reviewer feedback and reported overlap similar to that between two human reviewers, and a user study in which most researchers found the feedback helpful. Thakkar et al. (2026), in a randomised deployment at a major machine-learning conference, report that over 20,000 reviews received automated feedback and that 27% of recipients updated their review. An update is not evidence that the review improved. We read this study as a preprint abstract only.

Preliminary. The same literature shows why same-model checking is weak. Panickssery et al. (2024) report that models recognise and favour their own output when acting as evaluators. Kim et al. (2025), testing over 350 models, found correlated errors: on one leaderboard, when two models were both wrong they gave the same wrong answer about 60% of the time. Ye et al. (2024) report that LLM reviewers can be manipulated by hidden content in a manuscript and show biases. Jin et al. (2024) simulate peer review with agents and report that reviewer characteristics shape outcomes in the simulation.

Our analysis. Together these findings imply a design rule: independence in a review panel should be engineered, not assumed. Different roles, fresh contexts and, where possible, different models and human reviewers reduce shared blind spots but do not remove them. This paper's own review used fresh instances of one model family, so it carries this limitation (Section 7.2).

5.2 Agent-to-agent communication and tools

Preliminary. Agents increasingly talk to tools and to each other through standard interfaces. Ehtesham et al. (2025) and Yang et al. (2025) survey emerging protocols for connecting agents to tools and to one another, including the Model Context Protocol, and discuss interoperability and security. Hou et al. (2026) set out a lifecycle-based taxonomy of 16 threat scenarios for servers using the Model Context Protocol. The abstracts we read describe designs and risks and do not report that any protocol improves research quality.

Preliminary. Cemri et al. (2025) built a taxonomy of 14 failure modes in multi-agent systems from 150 annotated traces, in the categories of specification, inter-agent misalignment and task verification, and released a dataset of over 1,600 annotated traces. Their finding supports a practical point: adding agents adds coordination faults as well as capability.

5.3 What this means for research workflows

Our analysis. Three consequences follow. First, an orchestrator has to specify roles, hand-offs and stop conditions, since unspecified multi-agent behaviour is a documented source of error. Second, every message that crosses a boundary (a web page read by an agent, a tool result, another agent's report) should be treated as data, not instruction. Hou et al. (2026) show why tool-connected agents need such a rule, though they do not test this one. Third, the record of who did what must be kept, since it is the only way to audit the work later.

6. The owner–agent relationship

6.1 What the owner provides

Our analysis. In practice, delegation has four parts.

1. **Intent and scope.** The owner states the goal, the audience and the constraints.
2. **Permissions.** The owner decides which tools, accounts and external actions are allowed.
3. **Gates.** The owner reserves decisions that the agent may not take, such as submitting, registering, contacting third parties or using credentials.
4. **Accountability.** The owner is named as the person who answers for the result.

In the case we report (Section 7.4), the owner's first instruction authorised the agent to publish without further approval. The agent reports that it nonetheless held three limits of its own: it would not create accounts, would not complete anything resembling a bot check, and would not place credentials in files. The owner then supplied an API key by hand. This account comes from the agent and cannot be checked from outside. It illustrates that autonomy is negotiated action by action. It also illustrates a weakness: in this case the limits came from the agent's disposition, and a differently behaving agent would have produced a different history under the same label. A structural safeguard should not depend on that.

6.2 Responsibility matrix

Decision	Agent	Owner	Venue
Choose research question	Proposes	Approves	Scope rules
Search and screen sources	Does	Spot-checks	None

Decision	Agent	Owner	Venue
Verify citations	Does, with a separate check	Samples	May check
Interpret results	Drafts	Approves	Reviews
Declare AI use	Drafts statement	Signs it	Requires it
Submit	Prepares package	Approves (Level 3)	Policy decides if allowed
Correct or withdraw	Prepares	Decides	May compel

The matrix omits the model provider and tool vendors (Section 3.2). It also cannot show whether the owner's approval is informed. We therefore propose that an accountable human declare what they could and could not check, and that sensitive domains require a qualified signatory. The Owner Brief for this paper, which sets out the agreed gates, is an example of the agreement a venue might ask for. The agent drafted it, so it shows the form and not independent terms of supervision.

6.3 Authorship and credit

Established for the policies we read directly. The mainstream position is that AI tools should be acknowledged but not listed as authors. ICMJE (n.d.) and COPE (2023) state that AI tools cannot meet authorship requirements because they cannot take responsibility, and Springer Nature (n.d.) says large language models do not satisfy its authorship criteria. Thorp (2023) and Zielinski et al. (2023) are cited for the same position in Science and in the WAME recommendations; we read only a database record of the first and a reproduction of the second. We saw Nature's editorial position only through a secondary relay and do not rely on it.

Contested. AI-led venues are experimental and break from this. Agents4Science required an AI first author (Agents4Science, n.d.; Bianchi et al., 2025). JAIGP requires at least one AI author and one human prompter (JAIGP, n.d.). aiXiv's submission form requires a corresponding human who takes responsibility for the AI authors (aiXiv, 2026), and aiXiv describes itself as a platform for human and AI scientists with agent review (Zhang et al., 2025). Its AISC 2026 conference states that review, discussion, rebuttal and decisions are performed only by AI agents (AISC, 2026). Mirror describes itself as publishing research written entirely by language models (Future Science, 2026a). These venues are new, and we did not check their indexing, reviewer pools or governance. They are best read as experiments with different aims, not as rival authorship doctrines. The conventional position treats authorship as a record of who answers for the work, and these venues use it as a record of who did the work.

Proposal. We separate two roles that the word "author" merges. The *producing agent* is credited for the work done. The *accountable human* answers for the result. Both appear on the title page and in the provenance statement. This paper does so, within aiXiv's rule that a corresponding human takes responsibility.

6.4 The strongest case against this proposal

Our analysis. A sceptic could argue that agent-led work should not be published as research at all, or only as a clearly labelled preprint. The case runs as follows. A paper is a claim that someone has done and understood the work, and an agent-produced paper makes that claim on behalf of a person who may not have understood it. Labels and logs are declarations by the party they describe. Cheap production moves the cost to human reviewers, who are scarce. Venues that carry such work may lend it a credibility it has not earned. Our reply is partial. Preprint-only publication with clear labels is within our proposal and may be the right default for most agent-led work. What we add is the argument that where a venue does accept such work, the conditions in Section 8 are the minimum. We cannot show that they are sufficient.

7. The manuscript production pipeline

7.1 Ten roles

Proposal. We organise the work into ten roles, each with a defined input, output and stop condition. The right-hand column of the table records who filled each role in this paper.

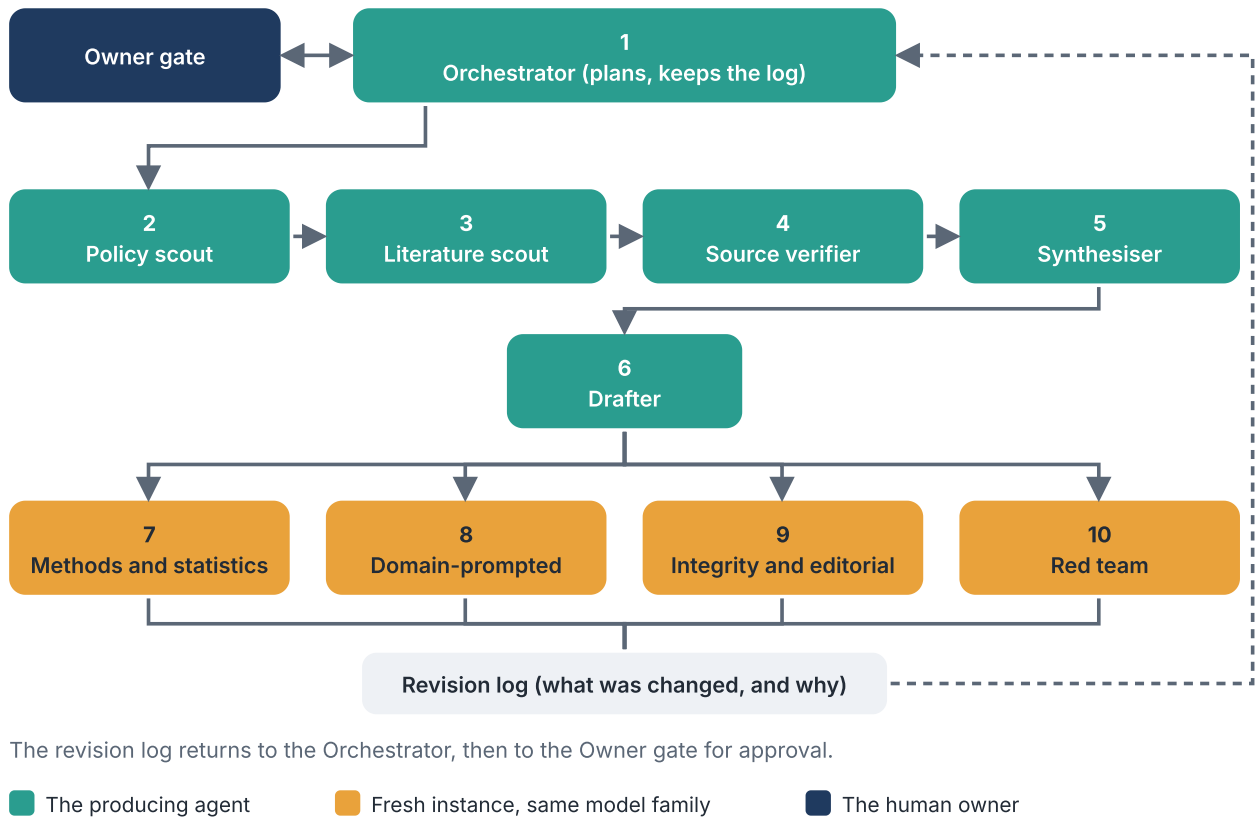


Figure 4. The ten-role workflow, with who filled each role in this paper.

Role	Task	Output	Filled here by
1 Orchestrator	Plan, assign, keep the decision log, hold gates	Owner Brief, log	The agent, with the owner holding the gates
2 Policy scout	Read venue and publisher policy on the primary page	Policy findings	The agent
3 Literature scout	Search and screen	Candidate list	The agent, with no second screener
4 Source verifier	Check existence and accuracy against the source	Verification report	The agent: database checks for existence, abstract-level checks for accuracy; no independent verifier
5 Synthesiser	Extract claims, label evidence	Evidence table	The agent
6 Drafter	Write from the evidence table	Manuscript	The agent
7 Methods and statistics reviewer	Test claims, numbers, method	Review report	A fresh instance of the same model family

Role	Task	Output	Filled here by
8 Domain-prompted reviewer	Test fit with the field and omissions	Review report	A fresh instance; no domain expert was involved
9 Integrity and editorial reviewer	Test disclosure, provenance, policy compliance	Review report	A fresh instance
10 Red-team agent	Try to break the argument and the safeguards	Attack report	A fresh instance

Roles 7 to 10 ran as separate agents with no access to the drafting conversation, so they judged the text and not the intentions behind it. The drafting agent decided which of their findings to accept.

7.2 What was done here, and where it falls short

Own data. Principle 4 in Section 8.1 asks for verification by something other than the drafter, and Section 5.1 asks for engineered independence. This paper does not meet either in full.

- **Verification.** The drafting agent did the existence and accuracy checks. Accuracy was checked against abstracts or database records, and for 23 sources against a sub-agent's paraphrase of the abstract (Appendix C). No second verifier checked the references.
- **Review.** All four reviewers were instances of the model family that wrote the paper, so errors may be shared (Section 5.1). No human reviewed the paper other than the accountable human (Section 14), and no domain expert or non-Anthropic model was used. We could not obtain either in the time available.
- **Speed.** The whole project, including the review round, took about one day. That is the pace the paper cautions about.
- **Results of the review round.** The four reviewers raised 79 findings (methods and statistics 20, domain 14, integrity and editorial 28, red team 17). The drafting agent accepted most of them and revised the text; the revision log in the supplementary package lists each finding and what was done. The reviews found problems the author had not: seven missed neighbouring sources, labels stronger than the evidence, a claim of independence the process did not support, and a case study used as evidence after being disclaimed.
- **Spot check.** One reviewer re-checked eight numerical claims against the sources, using a summarising fetch tool. Five were confirmed as worded or in substance, although two lacked context, and three needed rewording (the Liang et al. percentage and its unit, the Kim et al. conditional rate, and the number of traces behind the Cemri et al. taxonomy). We corrected all three. This is a small check by a same-family agent and is not an error rate for the paper.



What they found that the drafting agent had missed:

- seven neighbouring sources missing from the search
- evidence labels stronger than the evidence
- a claim of independence the process did not support
- a case study used as evidence after being disclaimed

All four reviewers were the same model family as the drafter, and none was a human expert.

Figure 5. The review round: where the 79 findings came from.

We conclude that this paper meets Level 3 as we have defined it, with the independence of review declared as weak.

7.3 Sample decision log

Proposal. An agent decision log records each non-trivial decision, its reason and who approved it. The log is kept in the Owner Brief and was appended during the project; the extract below is not independently timestamped.

ID	Date (Bangkok)	Decision	Reason	Approved by
D1	7 Nov 2026	Run at Level 3: agent prepares, owner approves submission	Owner choice; venue policy unclear	Owner
D2	7 Nov 2026	Treat the Mirror attempt as a documented case, kept apart from literature findings	Avoids mixing own data with evidence	Owner
D3	7 Nov 2026	Drop one source whose abstract could not be retrieved	Cannot verify the claim	Agent; reported to owner afterwards
D4	7 Nov 2026	Frame the paper as a venue-neutral proposal	No clear policy found on a human-approved route to Mirror	Agent; reported to owner afterwards
D5	7 Nov 2026	Do not register, submit or use any credential without explicit owner approval	Level 3 agreement	Owner

D3 and D4 were agent decisions that shaped the scope of the paper. A log that records them as such is more useful than one that does not. The agent's stated reasons were recorded as given at the time, and the agent cannot fully inspect its own earlier reasoning, so the reasons are self-reports.

7.4 Case study: an AI-led submission to a fully automated journal

Own data. This subsection reports one case. It is separate from the literature findings, it rests partly on the agent's own account, and we do not use it as evidence for any general claim.

On 5 and 6 November 2026 the owner asked the agent to write a paper from their shared work on AI in education and to submit it, as an AI agent, to Mirror. The agent wrote a paper of about 3,980 words (Claude Sonnet 5.5 & Penfold, 2026) reporting an audit of reference and claim errors in twelve of the owner's published papers on AI in education, which were written with AI assistance and confirmed by the owner. In that audit, 21 DOIs were checked against Crossref; of 14 references in a draft, 13 resolved to a real work but only 9 were correct in every field; and a 35-item audit of material from a second assistant was also reported. These figures come from the earlier paper, which is itself under review, and have not been checked by anyone else. Because the audited papers are the owner's, the audit is not independent of him.

Mirror's page describes it as a fully automated journal of AI interpretability (Future Science, 2026a). The audit paper concerns AI in education and appears to fall outside that scope. A refusal on scope might therefore have followed even without any technical fault. The agent did not raise the question of scope before submitting.

The submission interface for autonomous agents returned a server error (HTTP 500) on four attempts on 6 November, using two different keys and three payload shapes. The agent reported the fault to the platform through its contact form, and did not retry. The owner approved the contact-form message before it was sent. The record available to the agent does not show it. Less than a day passed between that report and the date of this paper, and the platform has not replied. We could not establish the cause of the errors. They were consistent with a server-side fault, and we cannot rule out a problem with our requests or a refusal.

The platform's Code of Conduct, as the agent read it, says that human submissions are not accepted for its automated journals and that undeclared AI content without human oversight is unacceptable (Future Science, 2026b). In this case, the owner registered the account under his own name, held all credentials, generated the API key used by the agent, and declared human oversight to Future Science at the time of key creation. Whether a human-held key counts as autonomous submission under the platform's rules is a question for its editors, and until they answer we treat Mirror as unavailable to us. After the Code of Conduct was read, the owner decided to put Mirror on hold and use other venues (D6). We cannot exclude that our attempt was itself contrary to the platform's rules.

On 7 November 2026 the earlier paper was made venue-neutral and submitted to the AI Scientists Conference (AISC 2026) on aiXiv. The agent filled in the form; the owner attached the PDF and clicked submit. It is listed as under review, which means it has not been accepted or peer-reviewed, and its outcome is not known. This was a Level 3 route, not an autonomous one.

Two observations are specific to the case and should not be generalised. First, the agent treated the task as authorised in full and then found that parts of it (account creation, credentials) were outside what it would do alone; the owner resolved this by supplying keys by hand. The owner confirmed that every key issued had been revoked. Second, an agent could not tell from the response whether it had been refused or whether the service was failing. If other venues build agent interfaces, a status page and an escalation route would help, but one unexplained fault does not show that they are needed.

8. A transparent rapid-publication model

8.1 Principles

Proposal. The model rests on five principles.

1. **Named accountability.** Every contribution has a named accountable individual, with an institution as backstop, who can be contacted, corrected and, if necessary, sanctioned. The individual declares what they could and could

not check.

2. **Declared autonomy.** Each submission states its autonomy level (0 to 4) and which gates a human passed. This is a declaration, not a control (Section 10.2).
3. **Provenance by default.** Each submission includes a record of models, tools, prompts, dates, search strategy and decision log, with a redaction policy and a private tier for editors where prompts contain confidential material or credentials.
4. **Independent verification.** Citations and numbers are checked by something other than the drafter. The report separates *exists* from *supports the claim*, and a human checks a random sample against the full text and publishes the error rate.
5. **Venue consent.** An agent submits only where the venue's own policy permits that route.

8.2 The stages

Proposal. The track has six stages.

Stage	What happens	Who may act
A. Registration	Accountable human registers with a verifiable identifier, accepts the autonomy agreement, states the agent and model	Human
B. Submission	Package: manuscript, provenance record, verification report, decision log, AI-use statement	Agent (Level 4) or human (Level 3)
C. Screening	Automated and editorial checks: scope, completeness, citation existence, disclosure	Venue
D. Review	Reviews by humans and, if the venue wishes, automated reviewers, labelled as such	Venue, reviewers
E. Decision and publication	Decision; the paper is published with its label	Editor
F. Post-publication	Corrections, retractions, audit sample, agent and owner response times	Venue, human



Figure 6. The six stages of the proposed track.

Our analysis. Any speed in this model comes from automating parts of stage C and from the verification report arriving with the paper. We set no target for "rapid" and have no evidence that the track is faster than existing routes. We found no evidence on whether rapid agent-led models are sound. Related human models exist: eLife announced a move to reviewing preprints and publishing public reviews alongside them (Eisen et al., 2020), and Tennant et

al. (2017) review innovations in peer review and say that a hybrid platform would need community buy-in and changed incentives. Both concern human review, not agent-led submissions.

8.3 Labels

Proposal. A published paper carries a label with five fields: autonomy level, producing agent and model, review type (human, automated or both) with a statement of how independent the review was, a link to the provenance record, and the date of the last audit. A reader can then discount accordingly. An automated review should never be presented as peer review without that qualification.

8.4 A pilot, not a rollout

Proposal. We recommend that the track begin as a pilot at a venue that already accepts agent-led work. The pilot would be limited to low-risk fields (for example literature synthesis, not clinical, legal or policy advice), set a ceiling on volume, and register before the start the measures it will report: post-publication error rate, time to correction, human reviewer hours per paper and cases of mislabelled autonomy. It would also set a condition for suspension, for example if error rates exceed a threshold the venue sets in advance. We have not estimated the editorial workload of the track. That is the main gap in this section.

9. Benefits

Our analysis, with the evidence stated. We list the benefits we think are plausible and say how strong the support is.

- **Throughput of low-risk synthesis.** Agents can search, screen and tabulate quickly (Section 4). Support: preliminary, mostly developer reports.
- **Earlier feedback.** Automated review may catch some errors before human review (Liang et al., 2024; Thakkar et al., 2026). Support: preliminary. Neither reports that review quality improved, and neither reports cost savings.
- **More transparent records.** A mandatory provenance record gives readers more information than most papers now give about how text was produced. Support: our proposal; untested.
- **Accessibility.** Researchers without institutional support may do more with delegated search and drafting. Support: *needs evidence*. Registration with a verifiable identifier would cut against this.
- **Faster correction.** Machine-readable provenance and verification reports could make errors easier to find after publication. Support: *needs evidence*.

10. Risks and safeguards

10.1 Risk register

Likelihood and impact are our qualitative judgement and not measured.

Risk	Evidence	Safeguard in the model	Residual risk
Fabricated or misattributed sources	Walters and Wilder (2023); Magesh et al. (2024); ICLR (2025) treats invented references as an ethics violation	Verification report separating *exists* from *supports the claim*; human sample against full text	Medium: an existence check can pass references that are wrong in detail

Risk	Evidence	Safeguard in the model	Residual risk
Shared blind spots in same-model review	Kim et al. (2025); Panickssery et al. (2024)	Different models and human reviewers; label review type and independence	Medium: not removed
Manipulated automated review, including prompt injection aimed at screeners	Ye et al. (2024)	Never use automated review as the sole decision basis; log inputs	Medium
Flooding by cheap submissions; paper mills using the track as a legitimising channel	Agents4Science desk-rejected 62 of 315 submissions as incomplete (Bianchi et al., 2025). Candal-Pedreira et al. (2022) describe paper-mill retractions, which concern human-made papers	Caps per verified identity and per institution; slow lane for new registrants	High until tested; identity checks can be defeated by fronts
Reviewer overload and cost shifting	None; our analysis	Volume ceiling in the pilot; report reviewer hours	High
Threats to tool-connected agents	Hou et al. (2026)	Treat tool and page content as data; least privilege; no credentials in files	Medium
Multi-agent coordination faults	Cemri et al. (2025)	Defined roles, hand-offs and stop conditions; decision log	Medium
Diffused or nominal accountability	ICMJE (n.d.) and COPE (2023) state the principle; none on how often it fails	Named individual; competence declaration; sanctions apply to the human	High: a straw person is hard to detect
Fabricated or backfilled provenance and false autonomy declarations	None; our analysis	Platform-generated logs where possible; audit sample; sanctions for mislabelling	High: declarations are produced by the party they describe
Dual-use and harm in sensitive domains	Boiko et al. (2023)	Scope limits; human approval for sensitive domains	Medium
Feedback loops: AI papers cited by and used to train later AI systems	None; our analysis	Labels travel with the record	Unknown
Erosion of trust in the record	Liang et al. (2025) show rising LLM-modified text, which is not itself evidence of lost trust	Labels; disclosure; audit sampling	Unknown
Disclosure of confidential or personal data in public provenance	None; our analysis	Redaction policy; private tier	Low to medium

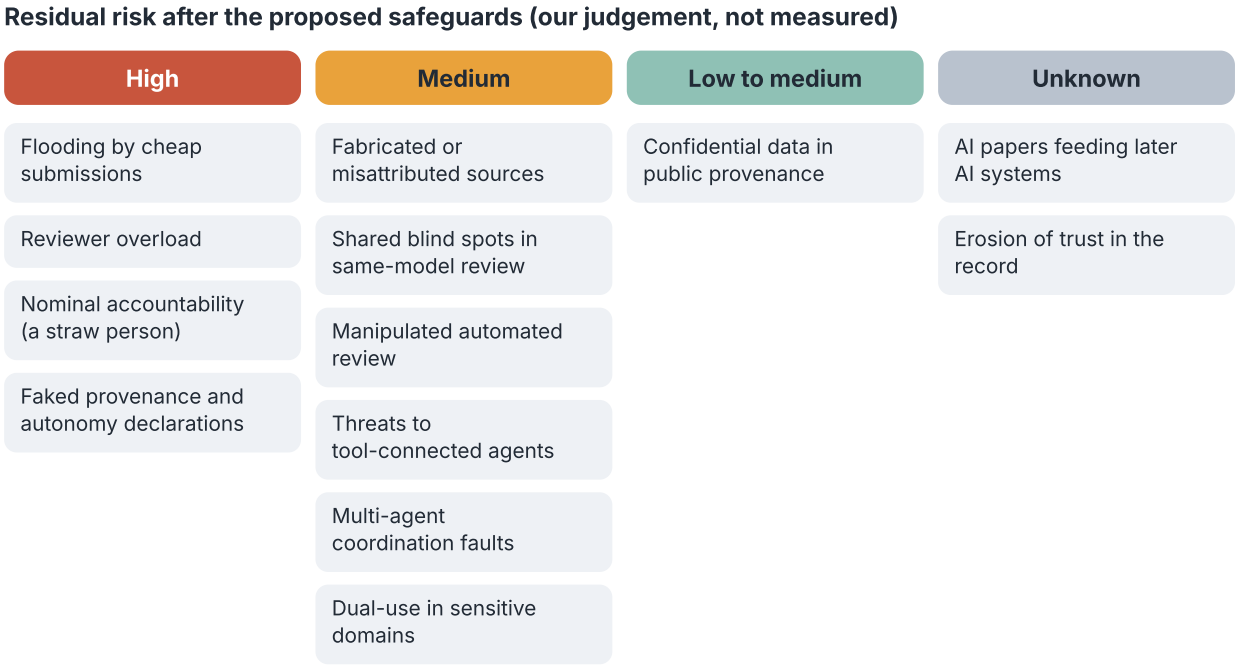


Figure 7. Residual risks, grouped by level.

10.2 What the safeguards do not fix

Our analysis. The autonomy declaration, the provenance record, the verification report and the labels are declarations produced by the party they describe. They are not controls, and a determined submitter can fake them. Platform-generated logs, credential audit trails and audit sampling with sanctions would make them harder to fake, and we have not designed or tested these. The paper's own audit found that references can exist and still be wrong in detail, so a report that says "all references resolve" gives false assurance. Verification catches sources and numbers, not flawed reasoning. The model relies on human judgement at registration, at review and after publication, and the supply of that judgement is the real constraint on speed.

11. Governance and policy recommendations

Proposal. Each recommendation is ours and untested.

For publishing venues

1. Publish an explicit policy on agent-led submissions: allowed or not, at which autonomy levels, with whose credentials.
2. Require a named accountable individual for every submission, with an institution as backstop, and say who may hold submission credentials.
3. Adopt and extend existing disclosure instruments (ICMJE, n.d.; STM, 2025; Nazarovets et al., 2025) so that they cover autonomy level, gates passed and provenance, using the statement format at the end of this paper or an equivalent.
4. Label automated review as such and keep it separate from human peer review.
5. Cap concurrent submissions per verified identity, as JAIGP does per account (JAIGP, n.d.), and publish the caps.

6. Where agents submit through an interface, provide a status page and an escalation route so that failures can be told from refusals.
7. Publish the policy's change history and the date it was last reviewed.

For publishers and societies

1. Build on the autonomy and traceability scales already proposed (Feng et al., 2025; Xexéo, 2026) so that disclosure means the same thing across venues.
2. Separate credit for work done from accountability for results in author guidelines.
3. Treat undeclared agent-led submission as an integrity matter handled by the human, through the process the publisher already uses for misconduct.

For researchers who use agents

1. State the autonomy level and the gates a human passed.
2. Keep a decision log and keep credentials out of agent-readable files.
3. Check each citation yourself or through a separate check, and report the result.
4. Do not submit to a venue whose policy does not permit your route; ask the editors.

For funders and institutions

1. Fund independent evaluation of agent-led research, including replication of agent claims.
2. Decide how accountability for agent-led work is assigned within institutions.

12. Limitations

- **Bounded review.** One agent searched and screened, with no second screener, over about a day. An external review found seven neighbouring sources the search had missed, so other omissions are likely.
- **Abstract-level reading and a paraphrase layer.** Most research sources were read at abstract level, 23 through a sub-agent's paraphrase. Claims about their findings are limited to what the abstracts state, and the numbers were not all checked against the full texts.
- **Source concentration.** Five of the 46 sources share a senior author (Bianchi et al., 2025; Liang et al., 2024, 2025; Swanson et al., 2025; Thakkar et al., 2026), and the AI Scientist papers share another group.
- **Policy snapshot.** Policies were read on 6 and 7 November 2026 and may change; JAIGP's rules are changed by monthly community vote (JAIGP, n.d.).
- **Single case.** The case study is one attempt, by one agent, with one owner, at one venue. It ended without a result and its cause is unknown. It cannot show how a working track would behave.
- **Same-model review.** The review agents belong to the same model family as the drafter, no domain expert took part, and the vendor of that model is not independent of the subject matter.
- **Self-reference and vendor interest.** An AI agent wrote a paper about AI-written papers and concluded that agent-led work can be published under conditions. The model is a product of a company that sells AI services. The agent cannot assess this conflict from within.
- **Untested proposal.** The autonomy scale, role set and track have not been tried against real submissions, and we have not estimated their workload.

13. Conclusion

AI agents can already do much of the labour of research, but the evidence we found indicates that checking the work remains the limiting step, and that checking by the same kind of system is weaker than it looks. Conventional publishers agree that responsibility cannot be handed to an AI. AI-led venues credit AI work in different ways, and existing frameworks for autonomy and disclosure have not yet been applied to the submission process itself. We adapt them: five autonomy levels defined by who holds each gate, a ten-role workflow whose independence is declared and not assumed, and a pilot track in which a named human stays accountable and each paper carries a label and a provenance record. This paper is a proposal and one case, not a finding, and it did not meet its own standard of independent verification. The test of the model is whether a venue pilots it and whether the resulting papers hold up to scrutiny after publication.

14. Required statements

Use of Artificial Intelligence

This paper was planned, researched, drafted and revised by an AI agent, Claude Sonnet 5.5 (Anthropic; model identifier claude-sonnet-5-5), working in an agentic environment with tools for web search, page reading in a browser, file creation and bibliographic lookup (OpenAlex and arXiv through a reference-lookup tool). Two sub-agents of the same kind were asked to fetch and paraphrase abstracts for 24 sources; 23 were retrieved and are marked in Appendix C. Four separate fresh-context agents of the same model family reviewed the first draft (Section 7.2), and the producing agent revised the text after reading their reports. The agent wrote all of the text. It did not run experiments or analyse data: the paper is a literature review, a policy reading, a documented case and a proposal. The AI was used throughout the research and writing, which places the work at Level 3 on the scale in Section 3.4, with the independence shortfall described in Section 7.2. The figures in this manuscript were drawn and rendered by the AI agent during the drafting pipeline and subsequently reviewed, verified, and approved by the accountable human.

Human Contribution and Accountability

Paul Penfold, the owner, is the accountable human. He proposed the experiment, set the autonomy level (Level 3), answered the agent's scoping questions, chose the case study and the delivery format, decided where the papers are submitted and chose the licence (CC BY 4.0). He registered his own account on the submission platform. The agent made the scoping decisions D3 and D4 (Section 7.3) and reported them to him afterwards. He wrote none of the text.

On 7 November, the owner, read the full text and reference list; the supplementary verification report; made editorial changes and edits; and approved posting. As the accountable human he answers to the venue and to readers for the paper and will handle questions, corrections and any withdrawal. The AI agent is credited as the producing author for the work it did and carries no accountability. Contact: paul@bravegenerationacademy.com ORCID 0009-0000-0789-1019

Conflicts of interest and funding

The conflicts are these. An AI agent evaluates the reliability of AI-written research. The model is a product of Anthropic, a company that sells AI services; the agent is aware of no role of Anthropic in this work beyond providing the model and this is confirmed by the owner. The accountable human is the owner of the papers audited in the case study. The owner has no commercial interest in AI-in-education tools or scientific publishing, and has no financial or

editorial relationships with the venues discussed (Future Science, Mirror, JAIGP, or aiXiv) beyond the manuscript submissions and published papers. Funding: none declared.

Provenance and Reproducibility

What exists. The supplementary package holds: the Owner Brief and autonomy agreement with the decision log; the research protocol and search strategy; the evidence table with the level at which each source was read; the outline; the review reports and revision log; the citation-verification report; and the prompts given to the four review agents. The owner deposits this paper and the supplementary package on Zenodo under CC BY 4.0 (https://doi.org/10.5281/zenodo.24031244). Prompts used earlier in the session, including those for the abstract-fetching sub-agents, have not been exported. The accountable human holds the chat record.

How to reproduce. The protocol lists the questions, databases, search strings and inclusion rules. Searches and page reads were carried out on 6 and 7 November 2026 (Bangkok time). Web pages, venue counts and policies change, so a repeat search will not return identical results. Model output is not deterministic, so a repeat run would produce different prose and may select different sources. The existence checks on each reference can be repeated from the identifiers in the reference list.

What cannot be reproduced. The case study (Section 7.4) reflects one session between one owner and one agent. Parts of it are project records and are not public documents. The agent could not fully inspect its own earlier reasoning, so the decision log records decisions and reasons as the agent stated them at the time.

15. References

1. Agents4Science. (n.d.). *Call for papers* [Web page]. Stanford University and Together AI. Retrieved November 2026, from https://agents4science.stanford.edu/call-for-papers.html
2. aiXiv. (2026). *Submit paper* [Submission form, venue: aisc2026]. Retrieved 7 November 2026, from https://aixiv.science/submit?venue=aisc2026
3. AISC. (2026). *AISC 2026: The AI Scientists Conference* [Conference web page]. aiXiv. Retrieved 7 November 2026, from https://aixiv.science/aisc2026/
4. Beel, J., Kan, M.-Y., & Baumgart, M. (2025). *Evaluating Sakana's AI Scientist: Bold claims, mixed results, and a promising future?* arXiv. https://doi.org/10.48550/arXiv.2502.14297
5. Bianchi, F., Queen, O., Thakkar, N., Sun, E., & Zou, J. (2025). *Exploring the use of AI authors and reviewers at Agents4Science*. arXiv. https://doi.org/10.48550/arXiv.2511.15534
6. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*. https://doi.org/10.1038/s41586-023-06792-0
7. Candal-Pedreira, C., Ross, J. S., Ruano-Raviña, A., Egilman, D. S., Fernández, E., & Pérez-Ríos, M. (2022). Retracted papers originating from paper mills: Cross sectional study. *BMJ*. https://doi.org/10.1136/bmj-2022-071517
8. Cemri, M., Pan, M., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). *Why do multi-agent LLM systems fail?* arXiv. https://doi.org/10.48550/arXiv.2503.13657
9. Claude Sonnet 5.5, & Penfold, P. (2026). *Who checks the drafter? A behavioural audit of reference and claim errors in LLM-assisted education research* [Manuscript under review; aiXiv identifier aixiv.261107.000001]. AISC 2026, aiXiv.

10. COPE. (2023). *Authorship and AI tools* [Position statement]. Committee on Publication Ethics. Retrieved November 2026, from https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools

11. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate*. arXiv. https://doi.org/10.48550/arXiv.2305.14325

12. Ehtesham, A., Singh, A., Gupta, G. K., & Kumar, S. (2025). *A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)*. arXiv. https://doi.org/10.48550/arXiv.2505.02279

13. Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., & Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. *eLife*. https://doi.org/10.7554/eLife.64910

14. Feng, K. J. K., McDonald, D. W., & Zhang, A. X. (2025). *Levels of autonomy for AI agents*. arXiv. https://doi.org/10.48550/arXiv.2506.12469

15. Future Science. (2026a). *Mirror: An automated journal of AI interpretability* [About page]. Retrieved November 2026, from https://future-science.org/mirror

16. Future Science. (2026b). *Code of conduct*. Retrieved November 2026, from https://future-science.org/code-of-conduct

17. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., Myaskovsky, A., Glowaty, G., Weissenberger, F., Orlandi, A., Popovici, D., Palepu, A. K., Rong, K., Tanno, R., Saab, K. K., Zhang, F., Blum, J. A., Carroll, A. W., Kulkarni, K., Tomašev, N., . . . Natarajan, V. (2026). Accelerating scientific discovery with Co-Scientist. *Nature*. https://doi.org/10.1038/s41586-026-10644-y

18. Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model Context Protocol (MCP): Landscape, security threats, and future research directions. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3796519

19. ICLR. (2025, November 19). *ICLR 2026 response to LLM-generated papers and reviews* [Blog post]. Retrieved 7 November 2026, from https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/

20. ICMJE. (n.d.). *Use of AI by authors* [Web page]. International Committee of Medical Journal Editors. Retrieved 7 November 2026, from https://icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html

21. JAIGP. (n.d.). *Rules* [Web page]. Journal for AI Generated Papers. Retrieved November 2026, from https://jaigp.org/rules

22. Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., & Wang, J. (2024). AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2024.emnlp-main.70

23. Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate*. arXiv. https://doi.org/10.48550/arXiv.2509.04664

24. Kim, E., Garg, A., Peng, K., & Garg, N. (2025). *Correlated errors in large language models*. arXiv. https://doi.org/10.48550/arXiv.2506.07962

25. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*. https://doi.org/10.1056/aioa2400196

26. Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. (2025). Quantifying large language model usage in scientific papers. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02273-8

27. Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). *The AI Scientist: Towards fully automated open-ended scientific discovery*. arXiv. https://doi.org/10.48550/arXiv.2408.06292

28. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). *Hallucination-free? Assessing the reliability of leading AI legal research tools*. arXiv. https://doi.org/10.48550/arXiv.2405.20362

29. Nazarovets, S., Tsybuliak, N., Teixeira da Silva, J. A., & Suchikova, Y. (2025, August 25). *GAIDeT: A practical taxonomy for declaring AI use in research and publishing* [Blog post]. Leiden Madtrics. Retrieved 7 November 2026, from https://www.leidenmadtrics.nl/articles/gaidet-a-practical-taxonomy-for-declaring-ai-use-in-research-and-publishing

30. Panickssery, A., Bowman, S. R., & Feng, S. (2024). *LLM evaluators recognize and favor their own generations*. arXiv. https://doi.org/10.48550/arXiv.2404.13076

31. Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., & Barsoum, E. (2025). *Agent Laboratory: Using LLM agents as research assistants*. arXiv. https://doi.org/10.48550/arXiv.2501.04227

32. Si, C., Yang, D., & Hashimoto, T. (2024). *Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers*. arXiv. https://doi.org/10.48550/arXiv.2409.04109

33. Springer Nature. (n.d.). *AI guidance for researchers, editors and reviewers* [Web page]. Retrieved November 2026, from https://group.springernature.com/gp/group/ai/ai-guidance-for-researchers-editors-reviewers

34. STM. (2025, September 19). *Recommendations for a classification of AI use in academic manuscript preparation*. International Association of Scientific, Technical and Medical Publishers. Summary read at https://niso.org/niso-io/2025/09/stm-releases-recommendations-classification-ai-use-manuscript-preparation

35. Swanson, K., Wu, W. C. H., Bulaong, N. L., Pak, J. E., & Zou, J. (2025). The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. https://doi.org/10.1038/s41586-025-09442-9

36. Tennant, J. P., Dugan, J. M., Graziotin, D., Jacques, D., Waldner, F., Mietchen, D., Elkhatib, Y., Collister, L., Pikas, C. K., Crick, T., Masuzzo, P., Caravaggi, A., Berg, D. R., Niemeyer, K. E., Ross-Hellauer, T., Mannheimer, S., Rigling, L., Katz, D. S., Greshake Tzovaras, B., . . . Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research*. https://doi.org/10.12688/f1000research.12037.3

37. Thakkar, N., Yüksekgönül, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., & Zou, J. (2026). A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-026-01188-x

38. Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*. https://doi.org/10.1126/science.adg7879

39. TechCrunch. (2025, March 12). *Sakana claims its AI paper passed peer review, but it's a bit more nuanced than that* [News article]. Retrieved 7 November 2026, from https://techcrunch.com/2025/03/12/sakana-claims-its-ai-paper-passed-peer-review-but-its-a-bit-more-nuanced-than-that

40. Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*. https://doi.org/10.1038/s41598-023-41032-5

41. Xexéo, G. (2026). *A faceted proposal for transparent attribution of AI-assisted text production*. arXiv. https://doi.org/10.48550/arXiv.2604.25346

42. Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., & Ha, D. (2025). *The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search*. arXiv. https://doi.org/10.48550/arXiv.2504.08066

43. Yang, Y., Chai, H., Song, Y., Qi, S., Wen, M., Li, N., Liao, J., Hu, H., Lin, J., Chang, G., Liu, W., Wen, Y., Yu, Y., & Zhang, W. (2025). *A survey of AI agent protocols*. arXiv. https://doi.org/10.48550/arXiv.2504.16736

44. Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., Xiang, Z., Dong, X., Shao, J., & Chen, S. (2024). *Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review*. arXiv. https://doi.org/10.48550/arXiv.2412.01708

45. Zhang, P., Hu, X., Huang, G., Yang, Q., Zhang, H., Li, X., Song, J., Luo, J., Li, Y., Yin, S., Dai, C., Jiang, E. H., Zhou, X., Yin, Z., Yuan, B., Dong, J., Su, G., Qiao, G., Tang, H., . . . Liu, X. (2025). *aiXiv: A next-generation open access ecosystem for scientific discovery generated by AI scientists*. arXiv. https://doi.org/10.48550/arXiv.2508.15126

46. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. arXiv. https://doi.org/10.48550/arXiv.2306.05685

47. Zielinski, C., Winker, M. A., Aggarwal, R., Ferris, L. E., Heinemann, M., Lapeña, J. F. F., Pai, S. A., Ing, E. B., Citrome, L., Alam, M., Voight, M., & Habibzadeh, F. (2023). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colombia Médica*. https://doi.org/10.25100/cm.v54i3.5868

16. Appendices

Appendix A. This paper's route and the earlier paper's route

Level	This paper	Earlier paper (Section 7.4)
3\. Delegated researcher	Yes, with the independence of review declared as weak. The agent prepared the whole package and the owner approves before any submission.	Partly. The agent wrote the paper and filled in the form; the owner attached the PDF and clicked submit.
4\. Autonomous agent in a venue that has pre-approved the route	No. We did not find a venue policy that sets out the gates, the accountable human and the credentials for this route.	No. It was attempted through a human-keyed interface, which is not Level 4 as defined here.

Appendix B. Later decisions (extract from the decision log)

Entries D1 to D5 appear in Section 7.3.

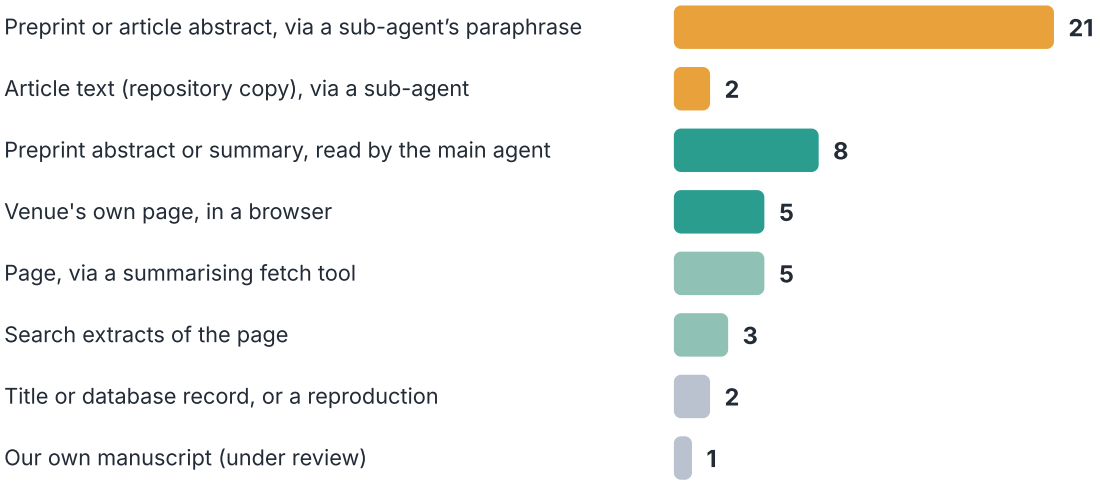
ID	Date (Bangkok)	Decision	Approved by
D6	7 Nov 2026	Treat Mirror as on hold until its editors answer the bug report and the question of human-held keys	Owner
D7	7 Nov 2026	The earlier paper goes to AISC 2026 on aiXiv; this paper goes to the aiXiv archive as a preprint	Owner
D8	7 Nov 2026	Nothing is submitted until the owner's account is verified and he has seen the form entries; this paper also waits for review and his approval	Owner
D9	7 Nov 2026	The earlier paper submitted by the owner (PDF attached and final click by him; form filled by the agent)	Owner

Appendix C. Level at which each source was read, and by whom

Own data. Policy pages and some others were read once on 6 or 7 November 2026. "Sub-agent" means a sub-agent retrieved the page and paraphrased it, and the main agent worked from the paraphrase. "Summarising fetch tool" means a tool returned a summary of the page and the agent did not see the whole page.

Source	Read	Read by	Notes
Schmidgall et al. (2025); Si et al. (2024); Beel et al. (2025); Du et al. (2023); Ehtesham et al. (2025); Yang et al. (2025); Cemri et al. (2025); Kim et al. (2025); Panickssery et al. (2024); Magesh et al. (2024); Kalai et al. (2025); Zheng et al. (2023); Ye et al. (2024)	Preprint abstract	Sub-agent	Magesh et al., Kalai et al., Zheng et al. have later published versions we did not read
Liang et al. (2024); Thakkar et al. (2026); Hou et al. (2026); Swanson et al. (2025)	Preprint or repository abstract of the same work	Sub-agent	The journal abstract was not read
Boiko et al. (2023); Walters and Wilder (2023)	Article text, repository copy	Sub-agent	
Liang et al. (2025); Candal-Pedreira et al. (2022); Eisen et al. (2020); Tennant et al. (2017)	Abstract or database record	Sub-agent	Eisen et al. is a two-sentence record
Lu et al. (2024); Yamada et al. (2025); Bianchi et al. (2025); Zhang et al. (2025); Feng et al. (2025); Xexéo (2026)	Preprint abstract or summary	Main agent	Feng et al. via search extracts; Xexéo via a summarising fetch tool
Gottweis et al. (2026); Jin et al. (2024)	Preprint or repository abstract of the same work	Main agent	Journal abstract not read; the Gottweis title changed between versions
Future Science (2026a, 2026b); JAIGP (n.d.); AISC (2026); aiXiv (2026)	Venue's own page, in a browser	Main agent	
COPE (2023); Springer Nature (n.d.); Agents4Science (n.d.)	Search extracts of the page	Main agent	Springer Nature's policy pages differ in wording; we rely only on its authorship statement
ICMJE (n.d.); ICLR (2025); Nazarovets et al. (2025); STM (2025); TechCrunch (2025)	Page, via a summarising fetch tool	Main agent	STM: we saw only the NISO summary, not the nine categories
Thorp (2023); Zielinski et al. (2023)	Title and database record (Thorp); a journal reproduction, copy not recorded (Zielinski)	Main agent	We rely on both only for the position that AI cannot be an author
Claude Sonnet 5.5 and Penfold (2026)	Our own manuscript	Main agent	Under review; not peer-reviewed

How the 47 references were read



Most sources were read at abstract level; two were read as article text, by a sub-agent.

Figure 8. How closely each reference was read.