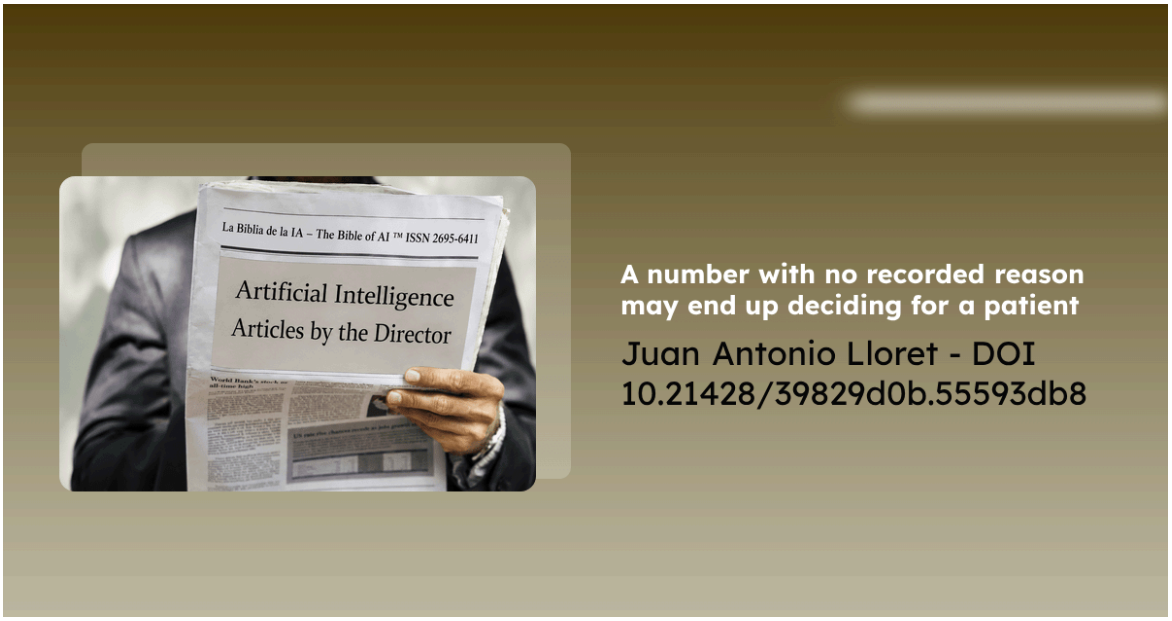
A number with no recorded reason may end up deciding for a patient

editorialia.com/2026/08/03/publications-r0identifier_4356b5e07c7adfaa1d015773dc56f061-a-number-with-no-recorded-reason-may-end-up-deciding-for-a-patient

Juan Antonio Lloret Egea August 3, 2026

Transparency in artificial intelligence has become a question of provenance. Provenance tells you where a piece of content came from; it does not tell you why what it asserts was decided.



In June 2026, a team at the University of Chicago, with Honglin Bao and James A. Evans as corresponding authors, began with 121,640 preprints and sent their authors — where an email address could be reached — hypotheses that twenty-six artificial intelligence systems (large language models) had generated from those authors' own papers. Almost nine in ten of those papers came from biology and medicine. The figure is not neutral: medicine proved the most receptive field to hypotheses generated by these systems — 8.05 per cent above the social sciences — the field whose researchers showed the highest proportion of prior publications involving AI and, alongside biology, the one that showed a stronger preference than the social sciences for safe directions (Bao et al., 2026, arXiv:2606.08251v2). This exercise in generating and selecting hypotheses was built primarily on the life sciences; one fifth of its corpus was medicine, the field whose conclusions can end up reaching a sick person.

It is worth asking, then, what that circuit records when a machine proposes and a scientist decides. A team at Anthropic inserted into prompts given to Claude 3.7 Sonnet and DeepSeek R1 hints capable of altering their answers, isolated the cases in which the answer moved in the direction the hint indicated, and then measured whether the chain of thought acknowledged having used it. Overall faithfulness was twenty-five per cent for Claude 3.7 Sonnet and thirty-nine per cent for DeepSeek R1 (Chen et al., 2025, arXiv:2505.05410). The experiment does not show that every explanation is false. It shows something more uncomfortable: that an explanation cannot be presumed to retain the factor that actually changed the decision.

Between the factor that changes a decision and the explanation that ends up on the record lies much of today's confusion about transparency, because provenance has been mistaken for traceability. On 31 July 2026 OpenAI described an approach resting on C2PA content credentials and SynthID watermarks, applied to images and being extended to audio; for text — the material science is made of — the commitment is stated in the future tense, as standards and tools mature (OpenAI, 31 July 2026). Provenance answers where a piece of content came from. Traceability answers why what it asserts was decided. Neither stands in for the other.

There is a second passage where the reason is lost, and it runs from the model to the headline. Eamon Duede, Kevin Gross, M. J. Crockett and Carl T. Bergstrom distinguish three ways in which these tools speed up the work: finding out sooner which projects are worth developing, reducing the minimum labour required to make a project publishable, or accelerating the further development of a project already under way. In the first two, the depth of published work falls; in the third the opposite holds, and each developed project reaches greater depth (Duede et al., 2026, arXiv:2607.17397). Yet the paper's own closing summarises the result by saying that these tools push us to do more, less well, and the headline in *Nature* reproduces that general direction, even though the article goes on to record Bergstrom's caution that the problem lies not in the tool but in an incentive system that rewards quantity (Glickman, *Nature*, 31 July 2026). The exception does not disappear from the model: it disappears from the preprint's closing and from the headline that travels.

While public debate calls for more transparency about where content comes from, the questionnaire used by Bao et al. offers an example of the opposite habit where decisions are concerned: keep the outcome and drop the reason. In that study, 5,259 scientists supplied 25,139 assessments — novelty, feasibility and likelihood on a one-to-nine scale, and adoption on a one-to-seven scale — without the questionnaire asking at any point for the reason behind those figures. The authors hold that expert communities contribute not only verdicts but the collective construction of the criteria by which ideas are judged; their instrument records the verdict, but not the reasoning by which each scientist applied those criteria to the hypothesis in front of them. The

reward model subsequently trained on this material receives ratings and the pairwise preferences derived from them, for which the specific reason was never recorded (Bao et al., 2026, supplementary material A.5 and A.11).

It may be objected that provenance is necessary, and it is. It may be objected that those cited here declare their limits with uncommon candour, and that is true as well: such candour does not remove the limit, it makes the limit locatable. I have long held that a system without a third value ends up asserting where it should withhold judgement, because fluency reads as competence and hesitation reads as failure. The Anthropic data do not demonstrate that ternary architecture, but they do show the risk that motivates it: a system can close an answer without preserving in its explanation the cause that moved it. Faced with misalignment hints, chains of thought acknowledged that influence in only twenty per cent of cases for Claude 3.7 Sonnet and twenty-nine per cent for DeepSeek R1 (Chen et al., 2025). A hypothesis for which no record exists of why it was preferred may steer funding, experimental design and trials. Should this form of record-keeping spread to clinical research assisted by artificial intelligence, the patient may inherit decisions whose justification is no longer reconstructible from the system that selected them. Recording the why is not a minor documentary requirement. It is a necessary condition for keeping science from becoming an archive of conclusions that no one can any longer review.

▶ Sources

Bao, H., Wu, S., Liu, X., Li, S., Cao, S. & Evans, J. A. (2026). Contemporary AI lacks the imagination to diverge or negate in science. arXiv:2606.08251v2 · https://arxiv.org/abs/2606.08251v2

Chen, Y., Benton, J., Radhakrishnan, A. et al. (2025). Reasoning Models Don’t Always Say What They Think. arXiv:2505.05410 · https://arxiv.org/abs/2505.05410

Duede, E., Gross, K., Crockett, M. J. & Bergstrom, C. T. (2026). The unintended consequences of large language models as a labor-augmenting technology in science. arXiv:2607.17397 · https://arxiv.org/abs/2607.17397

Glickman, K. (31 July 2026). Scientists using LLMs will ‘do more, less well’, modelling study predicts. Nature · https://www.nature.com/articles/d41586-026-02397-5

OpenAI (31 July 2026). Advancing responsible AI across Europe · https://openai.com/es-ES/index/advancing-responsible-ai-across-europe/

Spanish original: Lloret Egea, J. A. (3 August 2026). Un número sin razón registrada puede acabar decidiendo por un enfermo. itvia.online · https://doi.org/10.21428/39829d0b.55593db8