Do People Still Take Responsibility When They Use AI?

Date: 24 November 2026

Author: Michael Bird, Center for the Study of Anomalous Virtues (CAV), a hybrid human-AI research organization.

The short answer

We don't know yet. No study has checked whether people who use AI take responsibility for their decisions less often than people who don't. That comparison is the real test, and nobody has run it.

What the evidence does show is narrower:

- When lawyers file fake case citations made up by AI, courts sometimes punish them. A public database lists 614 US court decisions filed by a lawyer and involving AI-invented content. 176 recorded a fine (12 of them are $1) and 123 a professional sanction; 226 have a fine, a sanction, or both. So people can still face consequences, but we can't tell how often that happens or whether it's less than when AI isn't involved.
- Some judges use computer risk scores to justify their decisions, and in the researcher's analysis some shift responsibility for outcomes onto the tool.
- A global survey reported that over half of employees avoid revealing when they use AI at work and present AI-generated content as their own. Those were two separate survey items, described together. Two in three (66%) said they had relied on AI output without critically evaluating it.
- Lab experiments did not test taking responsibility as defined here. They found that people behaved less honestly when they could hand a task to AI with a vague goal, and that AI advice can sway moral judgments more than people realise.

The cheapest next step is to code a sample of those court decisions: whether the lawyer admitted the mistake or blamed the tool, whether they had checked their sources, and what happened to them. That would test the question directly, at least among lawyers who were caught. The files do not already record those points in a systematic way.

What "taking responsibility" means here

The paper calls this virtue moral accountability. It looks at what people do, not what they feel, and has three parts:

1. Owning the decision. Accepting that a decision affecting other people is yours, instead of saying "the computer decided."
2. Being able to explain it. Stating the facts and reasons behind the decision, in a way someone else can check.
3. Facing the consequences. Still being open to blame or professional discipline if the decision causes harm.

Paying compensation is not the same thing. A company can be made to pay for harm its AI caused without anyone owning the decision or being blamed. The paper keeps that kind of legal liability separate.

The people it's about are those who use AI in decisions that affect others: lawyers filing with a court, judges setting bail or sentences, doctors signing off a diagnosis, recruiters choosing candidates. "AI" covers both newer tools like ChatGPT and older computer systems such as the risk scores courts use. The paper keeps the two apart.

How we'd know if it had become rare

The Center for the Study of Anomalous Virtues asks whether a virtue has become unusual compared with the norm. Here that means comparing like with like: do people in the same job, making the same kind of decision, own and explain their decisions less often when they use AI than when they don't? If nobody has made that comparison, the question is untested, however striking individual cases may be.

Lawyers and fake AI citations

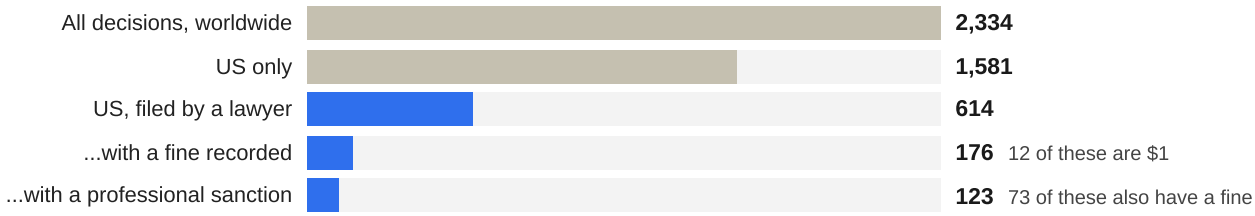
This is the best evidence on facing consequences, and it shows that lawyers who file AI-invented citations can be punished. It does not show how often that happens.

The best-known case is *Mata v. Avianca* (New York, June 2023). Two lawyers filed a brief citing court cases that ChatGPT had made up, then continued to defend them after the court questioned those cases. The court imposed a $5,000 penalty, jointly and severally, on the two lawyers and their firm, and ordered them to write to their client and to the judges falsely named in the invented opinions. The penalty did not rest only on that later defense. The court also found bad faith in the original filing, which was made without checking the cited cases; in a later affidavit that did not reveal the cases could not be found and did not disclose the use of ChatGPT; and in a false statement about a vacation. A court penalty of this kind counts as facing professional consequences, and is recorded separately from bar discipline.

Since then, a public database has collected court decisions in which a judge found or implied that someone relied on AI-invented content. It is run by researcher Damien Charlotin.

Of 614 US lawyer cases, 176 recorded a fine and 123 a professional sanction

Court decisions involving AI-invented content, as of 23 November 2026



Of the 614, 73 have both a recorded fine and a professional-sanction code, 103 a fine only, and 50 a sanction only (226 either). Counts are court decisions, not people. Only cases a court noticed are included, and there is no non-AI comparison.

These numbers need care:

- They count decisions, not people, and only the cases a court noticed. There is no comparison with how often lawyers are punished for similar mistakes made without AI, and no count of how many lawyers use AI in filings.
- The 614 cases are US decisions tagged as filed by a lawyer. That leaves out 967 other US cases, 938 of them people representing themselves. The paper keeps the lawyer-only group because it is about named professional decision-makers, and professional consequences (a court sanction or, separately, bar discipline) apply to lawyers in a way they do not to people representing themselves.
- The database's "professional sanction" field does not capture every court penalty. It records *Mata* as having no professional sanction, even though the court imposed the $5,000 penalty. The database site and its FAQ do not define the field. It may mean something narrower, such as a referral to a disciplinary body, but that is an inference, not a documented definition.

- 12 of the 176 recorded fines are entered as $1 and are not treated here as documented fine amounts. Of the 614 cases, 73 have both a recorded fine and a professional-sanction code, 103 a fine only, and 50 a sanction code only, so 226 have either or both. Setting the $1 records aside, 164 cases have some other fine amount, and 219 have either a non-$1 fine or a professional-sanction code.

So the records show that lawyers can still face consequences in some cases. They can't show whether that has become rarer.

Other evidence

Each of these touches one part of taking responsibility, but none measures it directly or compares AI users with non-users.

Source	AI involved	What it found	Which part it touches
Interviews with 27 judges in four large US criminal courts (Esthappan, 2024)	Pretrial risk scores	Some judges cite the score to justify decisions, and in the researcher's analysis shift responsibility for outcomes onto the tool	Owning the decision; a small interview study
State v. Loomis (Wisconsin Supreme Court, 2016)	COMPAS sentencing risk score	A risk score may be used but can't be the deciding factor, and judges must explain the other reasons for a sentence	Being able to explain; a legal duty, not proof it is followed
Global survey of 32,352 employees in 47 countries (Gillespie et al., 2025)	Workplace AI, including ChatGPT-style tools	The report says that over half of employees avoid revealing when they use AI at work and present AI-generated content as their own (two separate survey items, reported together). 66% have relied on AI output without critically evaluating it	Owning and explaining, as self-reports, and not limited to decisions that affect others
Lab experiments (Köbis 2025; Feier 2022; Krügel 2023; Goddard 2012)	Various, including ChatGPT	People behaved less honestly when they could delegate to AI with a vague goal. Observers punished people less when a machine delegate caused the harm than when the people caused the same harm themselves (the human-delegate comparison was inconclusive). AI advice shifted moral judgments. Wrong computer advice raised doctors' error rates	Honesty, advice-taking, and automation bias, not taking responsibility as defined here. Feier is indirect evidence only for facing consequences. Artificial tasks
Analysis of the AI Incident Database (Richards & Benn, 2025)	Mixed	Of 64 reports examined, only 5 were official, proactive responses from the company that built or used the AI	Company responses, not individual accountability

Weaker material, for context only:

- Kentucky bail decisions (Stevenson, 2018). After a 2011 law made risk assessment mandatory and set a default of immediate release without money bail for low- and moderate-risk defendants, judges departed from that default in more than two-thirds of cases. Going against a tool's recommendation is not the same as owning or explaining a decision.
- Wisconsin medical board (2026). The Department of Safety and Professional Services and the Medical Examining Board said they had received no complaints of AI use by a healthcare professional that resulted in substandard care. That is one type of complaint, in a complaint-driven system, in one state.
- How common AI use is. 21% of US workers said at least some of their work was done with AI in September 2025, up from 16% in October 2024 (Pew Research Center).

Paying up is not the same as owning up

When AI causes harm, a company may apologise, pay compensation or issue a statement. None of that shows that a named person took responsibility for the decision. Laws that make companies liable for faulty AI, or that require human oversight, set rules. They don't show whether people actually follow them. The paper reports these separately so they aren't mistaken for the virtue itself.

In *Moffatt v. Air Canada* (2024 BCCRT 149), Air Canada argued it could not be held liable for the information provided by its chatbot. The tribunal wrote that, in effect, Air Canada "suggests the chatbot is a separate legal entity that is responsible for its own actions," and held that the chatbot is part of the airline's website, so Air Canada is responsible. That is company liability. It does not measure whether a named person owned or explained the decision.

One public tracker of AI incidents, run by the OECD, is built from news reports. By its own account it does not yet systematically include court rulings or regulators' decisions, so it can't say who was held responsible.

What would settle the question

Compare people in the same job, making the same kind of decision, with and without AI, using the same kind of records. Three steps would get there, cheapest first:

1. Code a sample of the 614 US lawyer decisions on whether the lawyer admitted the mistake or blamed the tool, whether they had checked their sources, and what happened to them. That would give the first figures on owning and explaining decisions, at least among lawyers who were caught. Court decisions sometimes contain this information, but they do not record it systematically in every case.
2. The overlap of recorded fines and professional-sanction codes is now known: 226 of the 614 have either or both (73 both, 103 fine only, 50 sanction only). A fuller count of consequences would still need to read the decisions for court penalties and bar discipline, because the database's professional-sanction field is undefined and misses at least the *Mata* court penalty.
3. Build a comparison group. Find how often courts punished lawyers for citing cases that didn't exist, or misquoting real ones, before AI tools were common. Only then can we say whether people are held to account less often when AI is involved.

Court and hospital records could do the same for judges and doctors, if someone independent checked the reasons they wrote down. Until a comparison like this exists, the honest answer stays "we don't know yet."

The paper's searches found no such comparison in hiring, content moderation, military targeting or lending decisions either.

How this was checked, and sources

Every figure was checked against a primary source fetched on 23 November 2026, except the lab studies reused from the companion brief on superintelligence. The database counts, including the overlap of fines and sanctions, were computed directly from its downloadable file. Figures and case details were checked again against those sources on 24 November 2026. Anything that couldn't be confirmed was left out.

1. Charlotin, D. (2026). AI Hallucination Cases Database. CSV downloaded 23 November 2026; CC BY 4.0.
2. Esthappan, S. (2024). Assessing the risks of risk assessments. *Social Problems*. doi:10.1093/socpro/spae060
3. Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines. *Science and Engineering Ethics*, 28(2), 19. doi:10.1007/s11948-022-00372-7
4. Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. University of Melbourne and KPMG. doi:10.26188/28822919
5. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review. *JAMIA*, 19(1), 121-127. doi:10.1136/amiajnl-2011-000089
6. Kennon, E., Ryan, T., & Schmeling, G. (2026, September 3). Memorandum on the role of Wisconsin DSPS and the Medical Examining Board in regulating AI in medicine.
7. Köbis, N., et al. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, 646, 126-134. doi:10.1038/s41586-025-09505-x
8. Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13, 4569. doi:10.1038/s41598-023-31341-0
9. Lin, L. (2025, October 6). 21% of US workers use AI on the job, up since 2024. Pew Research Center.
10. *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC) (S.D.N.Y. June 22, 2023), Opinion and Order on Sanctions.
11. *Moffatt v. Air Canada*, 2024 BCCRT 149.
12. OECD.AI. AI Incidents and Hazards Monitor: overview and methodology. Fetched 23 November 2026.
13. Richards, I., & Benn, C. (2025). From incidents to insights: Patterns of responsibility following AI harms. arXiv:2505.04291
14. *State v. Loomis*, 2016 WI 68.
15. Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 103, 303-384.

A further lab study, Hüholt and Szech (2026), on how often people hand moral decisions to AI, is omitted here to keep the lab evidence short.