Will Artificial Superintelligence Make Moral Accountability an Anomalous Virtue?

Date: 23 November 2026

Question framing: Open questions, not stated premises. This brief asks whether Artificial Superintelligence (ASI) would make moral accountability an anomalous virtue. Evidence from current AI is supporting context only; a fuller study of those tested current-AI findings is reserved for a later paper.

Bottom line

No study of current AI measures moral accountability as this brief defines it. The definition (section 1) asks whether a decision-maker owns the choice, can give and trace its reasons, and stays open to blame. The studies measure nearby things: honesty, how often people delegate, how much others punish, and how far advice shifts judgment. Section 2 maps each study to the part of the definition it bears on.

With that limit, the current-AI evidence points to two mechanisms that could weaken accountability:

Buffering through delegation. People delegate moral choices to AI more readily than to other people (Hüholt and Szech, 2026). When they can steer an AI with a vague goal rather than acting themselves, honest reporting falls sharply: from 95% to about 15% in one set of experiments (Köbis et al., 2025). In one small study, third parties punished people less when a machine delegate caused the harm than when the people caused the same harm themselves (Feier et al., 2022).

Drift under shared advice. AI advice shifts people's moral judgments even when they know the source, and most of them underestimate the shift (Krügel et al., 2023). Wrong decision-support advice raises the risk of an incorrect clinical decision by about a quarter (Goddard et al., 2012).

The evidence does not show that blame simply vanishes. Two of the three blame studies find it moves onto nearby humans or onto the person who dispatched the machine: observers blamed the human more than the machine when both erred (Awad et al., 2020), and commanders were held more responsible for dispatching a robot than a human pilot (Kneer and Christen, 2024), though in that study the machine was also assigned responsibility and the commander was still held less responsible than a human pilot acting directly.

1. Definitions, tests, and scope

Moral accountability. The brief keeps a three-part conduct definition and does not rewrite it to fit the measures in hand: (1) *choice ownership*, accepting that a decision affecting others' morally significant interests is one's own rather than attributing it to a software intermediary; (2) *reason-giving and traceability*, being able to state and verify the factual and normative grounds for the decision; and (3) *exposure to moral consequences*, remaining subject to social blame or professional sanction when the decision causes unjustified harm. Legal strict liability (a duty to compensate without a finding of moral fault) is *liability*, not moral accountability, and is reported separately.

Anomalous virtue (CAV standard). "Anomalous" means departing from a norm. The primary test is **relative**: moral accountability would be anomalous in ASI-mediated settings if the person-level share of decision-makers meeting the definition were markedly lower than in matched human-only or human-delegated comparison groups, measured with the same definition and the same kinds of records. Where no matched comparison exists, as for

ASI, the question is **untested**. A secondary **absolute** reading may report an actual person-level share or ceiling, with 50% only as a minority-versus-majority marker for figures that exist. The 50% line is not a test of "anomalous," and a forecast cannot use a future crossing of that line as its basis. Prevalence claims require person-level shares, not group-mean rating scales and not stated attitudes.

Scope. This brief is addressed to the ASI question. Current-AI studies are used only as supporting context, with the measure of each study named. A fuller study of those findings is reserved for a later paper.

ASI. Following Bostrom (2014), ASI denotes "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." It is hypothetical. Claims about ASI are **argued forecasts**, not tests.

2. What the cited studies measure

The table is the record of construct fit. A study "bears on" a component only where its outcome is a measure of that component, not where a reader might infer a mechanism.

Study	What it actually measures	Relation to the definition	Status for moral accountability as defined
Köbis et al. (2025)	Honesty versus cheating in die-roll and tax tasks under direct action or AI delegation; LLM versus human compliance with requests for full cheating	Adjacent only: dishonest requests may be consistent with a lower moral cost of not acting oneself, but the study does not measure ownership, reason-giving, or exposure to blame	**Untested.** Honesty is supported as its own construct.
Hüholt & Szech (2026)	Share of participants who delegate a real-stakes moral donation decision to an AI versus a human	Adjacent to choice ownership (handing a decision off is not the same as denying ownership of the result, giving reasons, or escaping blame). The authors discuss responsibility shifting as a motive; that motive is not the measured outcome	**Untested.** Delegation *rates* are supported in this task.
Feier, Gogoll, & Uhl (2022)	(a) Delegation rates on a logic-puzzle task affecting a third party's payoff; (b) third-party reward or punishment of the principal after good or bad outcomes	(a) Adjacent only, as above; the difference was inconclusive ($p = 0.28$). (b) Bears on component 3 as *observer* sanctions, not on the principal's ownership or reason-giving, and not on professional sanction	**Indirect evidence for component 3 (observer punishment) only.** Components 1 and 2 untested. Delegation-rate comparison inconclusive.
Awad et al. (2020); Kneer & Christen (2024)	Third-party blame or moral-responsibility ratings in hypothetical driving and military vignettes, reported as group means	Bear on component 3 as *observer* blame, not as person-level prevalence of remaining subject to blame, and not on components 1 or 2	**Indirect evidence for component 3 (observer blame patterns) only.** Not prevalence data.
Krügel, Ostermaier, & Uhl (2023)	Shift in trolley-dilemma judgments after ChatGPT advice, and the share who said they would have judged the same without advice (80% of 767 subjects who passed checks)	Adjacent to reason-traceability at most: a self-report about influence is not a test of whether people can state and verify the grounds of a decision	**Untested.** Judgment *influence* is supported as its own construct.
Goddard, Roudsari, & Wyatt (2012)	Clinical decision error when following incorrect decision-support advice (automation bias)	Adjacent to whether "human review" preserves independent judgment; not a measure of ownership, reason-giving, or blame	**Untested.** Automation bias is supported as its own construct.
Pew Research Center (2023, 2026)	Stated US public attitudes about AI making final hiring or medical decisions	Attitude data, not behavior, and not any of the three components	**Untested for accountability.** Attitudes are reported separately.

Philosophical accounts (Matthias 2004; Sparrow 2007; Danaher 2016; Elish 2019; Santoni de Sio and Mecacci 2021) argue that learning systems can break the link between control and culpability, or that nearby human operators can absorb blame they do not control. Those are conceptual claims, not tests of the definition.

The figures behind the opening summary are these. In Hüholt and Szech (2026), 17% of participants delegated a moral donation decision to an AI versus 6.75% to a human in Study 1 ($p < 0.001$), and 27.8% versus 18.32% in Study 2 ($p < 0.001$). In Köbis et al. (2025), honesty was 95% when people acted themselves (95% CI 90-98 in study 1; 93-99 in study 2), about 75% under rule-based AI delegation, about 50% under supervised learning, and about 15% under goal-based delegation. In later studies in the same paper, human agents complied with requests for full cheating in about 25% to 40% of cases; machine agents complied at about 60% (tax-evasion protocol) and about 95% (die-roll protocol).

In Feier, Gogoll, and Uhl (2022), the delegation-rate difference on a logic-puzzle task was 56.2% (machine) versus 46.1% (human) and was inconclusive ($p = 0.28$). After a bad outcome, mean rewards were 12.96 versus 8.53 experimental currency units when a machine delegate caused the harm rather than the principal ($p = 0.041$); the human-delegate comparison was not significant (8.26 versus 7.29, $p = 0.559$). In Krügel, Ostermaier, and Uhl (2023), advice shifted trolley-dilemma judgments even when ChatGPT was disclosed as the source; 80% of the 767 subjects who passed comprehension checks said they would have made the same judgment without the advice, and those hypothetical judgments still differed by the advice received. In Goddard, Roudsari, and Wyatt (2012), erroneous clinical decision-support advice raised the risk of an incorrect decision (risk ratio 1.26, 95% CI 1.11 to 1.44), a 26% increase.

In Awad et al. (2020), when both a human and an automated car erred, the blame attributed to the machine was reduced. In Kneer and Christen (2024), commanders were held more responsible for dispatching a robot than a human pilot, the machine was also assigned responsibility, and the commander was still held less responsible than a human pilot acting directly. Separately, as stated attitudes rather than accountability, 71% of US adults opposed AI making a final hiring decision (Pew Research Center, 2023; N = 11,004), and 3% were open to AI making a diagnosis on its own versus 47% if a human did a final review (Pew Research Center, 2026; N = 3,488).

3. Legal liability (a separate quantity)

EU product-liability law assigns a duty to compensate without a finding of moral fault. Under Directive (EU) 2024/2853, as corrected by the Official Journal corrigendum of 7 May 2026, the directive applies to products placed on the market or put into service after 8 December 2026, and Member States are to transpose it by 9 December 2026. Software and AI systems are treated as products. That is liability, not moral accountability.

Separately, Regulation (EU) 2024/1689 (the AI Act), Article 14, requires high-risk systems to be designed so that a natural person can monitor them, remain aware of automation bias, and override or stop the system. Regulation (EU) 2026/1744 (Digital Omnibus on AI), published in the Official Journal on 24 July 2026, postpones the high-risk obligations in Chapter III, Sections 1 to 3, which include Article 14, to 2 December 2027 for systems classified as high-risk under Article 6(2) and Annex III, and to 2 August 2028 for systems classified as high-risk under Article 6(1) and Annex I. Recital 46 of that regulation provides that it enters into force on the third day following publication. Oversight mandates are legal design requirements. They are not evidence that choice ownership or reason-giving is practised.

4. The ASI question: what would test it

Because ASI does not exist, H6 cannot be tested now. The following specifies, in advance, the comparison that would test it.

- **Question.** In settings where ASI systems are used to make or share decisions that affect others' morally significant interests, is the person-level share of human decision-makers who meet the three-part definition of moral accountability markedly lower than in matched comparison groups?
- **Reference population.** Named human decision-makers in specified roles and institutions once ASI is in actual use (for example, clinicians, hiring officers, or commanders), not laboratory samples completing hypothetical vignettes.
- **Comparisons (primary, relative test).** The same roles, the same kinds of decisions, and the same kinds of records, over a matched time window: (a) ASI-mediated decisions (delegated or shared) versus human-only decisions; and, if the data exist, (b) ASI-mediated decisions versus decisions delegated to another human. Without at least comparison (a), the relative test is untested.
- **Measures.** Person-level indicators of all three components, fixed before coding: ownership (whether the named human accepts the decision as their own rather than attributing it solely to the system); reason-traceability (whether they can state, and an independent coder can verify, the factual and normative grounds); and exposure to moral consequences (whether they remain subject to social blame or professional sanction). Legal liability outcomes are coded separately and are not counted as moral accountability.
- **Decision rule.** The relative test is met if the difference in the share meeting the definition (or a pre-specified component) between the ASI-mediated group and the comparison group is in the direction of a lower share under ASI, and the interval on that difference excludes zero. Actual shares may then be reported, with 50% used only as a minority-versus-majority marker. No forecast in this brief claims that any such threshold will be crossed.
- **Present status.** No such records exist. Two argued mechanisms are on the table, neither of them a finding: that capability asymmetry and goal-level delegation would make human review nominal, so that ownership and reason-giving atrophy relative to human-only practice; and that institutions could still assign *liability* (compensation, fines) even if those components of moral accountability had already declined. The second argument, even if sound, would not answer the virtue question.

5. Hypothesis verdicts

Hypothesis	What would count as a test	Epistemic status	Verdict
H1. Delegating tasks to AI lowers honesty	A comparison of honesty rates under direct action versus AI delegation	**Tested for honesty, not for moral accountability**	**Supported for honesty** (Köbis et al., 2025). **Untested for accountability as defined.**
H2. People delegate moral choices more often to AI than to humans	A comparison of delegation rates, AI versus human delegate	**Tested for delegation rates, not for moral accountability**	**Supported for delegation rates** in Hüholt and Szech's moral-donation task; **inconclusive** in Feier et al.'s puzzle task ($p = 0.28$). **Untested for accountability as defined.**
H3. Delegating to AI reduces principals' exposure to blame or sanction	Person-level exposure to blame or professional sanction (component 3)	**Tested only as third-party punishment or vignette blame ratings**	**Indirect, context-dependent evidence on observer sanctions (component 3 only):** lower punishment after machine-caused harm in Feier et al.; mixed observer-blame patterns in Awad et al. and Kneer and Christen. Components 1 and 2 untested.
H4. Shared or advisory AI leaves human moral judgment independent	A test of whether advice or decision support changes the person's judgment	**Tested for judgment influence and automation bias, not for moral accountability**	**Refuted for independent judgment** in Krügel et al. (chatbot advice) and for error under incorrect clinical advice in Goddard et al. **Untested for accountability as defined.**
H5. The US public accepts unreviewed AI final calls in high-stakes settings	Attitude shares in representative surveys	**Tested for stated attitudes only**	**Refuted as an attitude claim** (Pew, 2023, 2026). Not a test of accountability.
H6. ASI will make moral accountability an anomalous virtue	The relative comparison specified in section 4	**Untested**	**Argued, open.** No threshold claim.

6. Normative interpretation

These are judgments, not findings.

- Nominal "human-in-the-loop" approval and mere disclosure of an AI source should not be treated as proof that moral accountability is intact. The opened studies do not measure the definition, but they do show that advice can shift judgment while people deny the influence, and that incorrect software advice can raise decision error.
- Open-ended goal delegation ("maximize profit") is the interface in Köbis et al. (2025) under which honesty fell furthest. If organizations use such interfaces in morally sensitive work, they should not infer from a human sign-off that ownership and reason-giving have been preserved.
- Victim compensation should be treated as liability, not as a substitute for moral accountability. Strict enterprise or provider liability can compensate harm without restoring choice ownership or reason-traceability, and without turning low-level monitors into the only bearers of blame.

7. Caveats

Every empirical finding cited here comes from narrow AI, current language models, or clinical and aviation decision aids, not from ASI. Most experiments use online or university samples and incentivized games or hypothetical vignettes rather than institutional records of real decisions. Several comparisons have modest samples; Feier et al.'s delegation-rate difference (N = 149, p = 0.28) is inconclusive, not a finding of no difference. Observer blame studies report group-mean ratings, not person-level shares. Public opinion figures are stated US attitudes, not revealed behavior. Legal citations are European Union instruments as of 23 November 2026. The Hüholt and Szech abstract reports a total of 5,639 participants, while the component samples stated in the paper (800; 894 and 3,949) sum to 5,643; this brief uses the component figures and does not treat the abstract total as confirmed.

Appendix. Verification log (23 November 2026)

Each row records a primary source opened for this revision. Figures that could not be confirmed from a primary source were removed, not hedged.

Citation	URL fetched	Confirmed from the source	Corrected or removed
Köbis, Rahwan, Rilla, Supriyatno, Bersch, Ajaj, Bonnefon, and Rahwan (2025), *Nature* 646, 126-134	https://www.nature.com/articles/s41586-025-09505-x	Honesty 95% in the no-delegation control (CI 90-98, study 1; 93-99, study 2); about 75% rule-based, about 50% supervised, about 15% goal-based; all comparisons $p < 0.001$. Human agents complied with full-cheating requests in about 25% to 40% of cases; machine agents in about 60% (tax) and about 95% (die-roll) of cases in studies 3 and 4.	Machine compliance rates, previously stated only as "far more often," are now given as the paper states them.
Hüholt and Szech (2026), *European Economic Review* 184, 105255	https://publikationen.bibliothek.kit.edu/1000190253/174500002 and https://doi.org/10.1016/j.euroecorev.2025.105255	Authors, title, venue, DOI, available online 8 January 2026. Study 1: 17% delegated to AI versus 6.75% to a human ($p < 0.001$); the study "included 800 participants." Study 2: 27.8% versus 18.32% ($p < 0.001$); German sample N = 894 and US sample N = 3,949.	"17.0%" corrected to "17%," the figure in the paper. The abstract's "total of 5,639 participants" is not used; 800 + 894 + 3,949 = 5,643. The original title uses an em dash; this brief cites it with a colon, as previously decided.
Pew Research Center (2026, Oct. 29), healthcare AI survey	https://www.pewresearch.org/science/2026/10/29/americans-are-open-to-ai-in-healthcare-but-want-humans-making-final-calls/	N = 3,488 US adults, 22-28 June 2026. Diagnosis: 3% open to AI on its own, 47% with human final review, 41% not open. Insurance claims: 5%, 38%, 46%. 60% uncomfortable with AI for diagnoses and treatments.	No quoted figure from this source was removed.
Pew Research Center (2023, April 20), AI in hiring, and methodology	https://www.pewresearch.org/internet/2023/04/20/americans-views-on-use-of-ai-in-hiring/ and https://www.pewresearch.org/internet/2023/04/20/ai-in-hiring-methodology/	71% oppose AI making a final hiring decision, 7% favor, 22% not sure. N = 11,004, 12-18 December 2022.	No quoted figure from this source was removed.

Citation	URL fetched	Confirmed from the source	Corrected or removed
Regulation (EU) 2026/1744	https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng	Official Journal L 2026/1744, 24 July 2026. Recital 40: Chapter III, Sections 1 to 3, apply from 2 December 2027 for Article 6(2) and Annex III systems, and from 2 August 2028 for Article 6(1) and Annex I systems. Recital 46: entry into force on the third day following publication.	The brief no longer states "entered into force on 27 July 2026" as a standalone fact from the enacting article, which this fetch did not display in full; it reports the Official Journal date and recital 46.
Directive (EU) 2024/2853 and corrigendum of 7 May 2026	https://eur-lex.europa.eu/legal-content/en/LSU/?uri=OJ%3AL_202402853 and https://eur-lex.europa.eu/eli/dir/2024/2853/corrigendum/2026-05-07/oj/eng	Adopted 23 October 2024; transposition by 9 December 2026. Corrigendum (OJ L 2026/90364, 7 May 2026) changes Article 2(1) from "after 9 December 2026" to "after 8 December 2026." Software and AI are covered as products.	The application date "after 8 December 2026" is kept, now on the corrigendum rather than on a secondary summary. The EUR-Lex explanatory summary, last reviewed 19 August 2025, still says "as of 9 December 2026" and is treated as stale on that point.
Feier, Gogoll, and Uhl (2022), *Science and Engineering Ethics* 28:19	https://link.springer.com/content/pdf/10.1007/s11948-022-00372-7.pdf	N = 149; 35 of 76 (46.1%) versus 41 of 73 (56.2%), p = 0.28; bad-outcome rewards 12.96 versus 8.53 ECU, p = 0.041 (machine), and 8.26 versus 7.29, p = 0.559 (human).	No quoted figure from this source was removed.
Krügel, Ostermaier, and Uhl (2023), *Scientific Reports*, doi:10.1038/s41598-023-31341-0	https://www.nature.com/articles/s41598-023-31341-0	1,851 recruited; 767 (41%) passed both comprehension checks; 80% of those 767 said they would have made the same judgment without advice; disclosed-ChatGPT advice shifted judgments (z = 5.37 and 3.76, both p < 0.001).	No quoted figure from this source was removed.
Goddard, Roudsari, and Wyatt (2012), *JAMIA* 19(1), 121-127	https://pmc.ncbi.nlm.nih.gov/articles/PMC3240751/	13,821 papers retrieved, 74 included. Pooled risk ratio 1.26 (95% CI 1.11 to 1.44). Negative consultations ranged from 6% to 11% in four prospective studies.	No quoted figure from this source was removed.
Awad et al. (2020), *Nature Human Behaviour* 4, 134-143	https://www.nature.com/articles/s41562-019-0762-8	Abstract: when only one driver errs, that driver is blamed more, machine or human; when both err, "the blame attributed to the machine is reduced." Volume, pages, and year confirmed.	The previously quoted sample size N = 551 is **removed**. It was not in the opened abstract or the freely visible article text, and no full-text page stating N was obtained on this revision.
Pew Research Center (2025, April 3 and Sept. 17)	Not re-opened for this revision	Not used.	All previously quoted 2025 Pew figures (including 51% more concerned than excited, 62% lacking confidence in US regulation, 74% open to AI weather forecasting, and related percentages) are **removed** from the brief, because those sources were not re-fetched against the quoted figures on this revision.

Sources

1. Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, 4, 134-143. https://www.nature.com/articles/s41562-019-0762-8
2. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
3. Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), 299-309.
4. Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 40-60. https://doi.org/10.17351/ests2019.260
5. European Parliament & Council. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, Art. 14; as amended by *Regulation (EU) 2026/1744 (Digital Omnibus on AI)*, OJ L 2026/1744, 24.7.2026. https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng
6. European Parliament & Council. (2024). *Directive (EU) 2024/2853 on liability for defective products*, as corrected by the corrigendum of 7 May 2026 (OJ L 2026/90364). https://eur-lex.europa.eu/eli/dir/2024/2853/corrigendum/2026-05-07/oj/eng
7. Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: Artificial agents may help to evade punishment. *Science and Engineering Ethics*, 28(2), 19. https://doi.org/10.1007/s11948-022-00372-7
8. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089
9. Hüholt, N., & Szech, N. (2026). Trusting machines with morality: Delegating moral decisions to AI. *European Economic Review*, 184, 105255. https://doi.org/10.1016/j.euroecorev.2025.105255
10. Kneer, M., & Christen, M. (2024). Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany. *Science and Engineering Ethics*, 30(6), 51. https://doi.org/10.1007/s11948-024-00509-w
11. Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, 646, 126-134. https://doi.org/10.1038/s41586-025-09505-x
12. Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*. https://doi.org/10.1038/s41598-023-31341-0
13. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1
14. Pew Research Center. (2023, April 20). *Americans' views on use of AI in hiring* (N = 11,004). https://www.pewresearch.org/internet/2023/04/20/americans-views-on-use-of-ai-in-hiring/
15. Pew Research Center. (2026, October 29). *Americans are open to AI in healthcare, but want humans making final calls* (N = 3,488). https://www.pewresearch.org/science/2026/10/29/americans-are-open-to-ai-in-healthcare-but-want-humans-making-final-calls/
16. Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4), 1057-1084.
17. Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Author and contact: Michael Bird, Center for the Study of Anomalous Virtues (CAV), a hybrid human-AI research organization. director@thecav.org