
Behavioural Indicators of Fault in Large Language Models¹

Martin Radzaj
Barrister
Founder, Lawyers for AI Safety

with
Apart Research

Abstract

As increasingly autonomous AI systems are deployed, they inevitably take consequential actions that may cause harm. Legal systems traditionally distinguish liability based on an actor's state of mind (knowledge, recklessness, or negligence) which is often inferred. We investigate whether this legal framework can be applied to Large Language Models (LLMs) to determine if harmful choices coherently track these epistemic states. We introduced a novel behavioral framework, testing three models over 3,240 API calls across four fictional scenarios. We systematically manipulated two variables, the model's knowledge of latent harm and its purpose, to isolate behavioral signs of fault. Our findings reveal that while knowledge of harm generally reduces harmful action, goal pressure produces unstable results, acting more as a driver of model activity than an indicator of intent. We identified significant "option-order" bias, where model choices were driven more by the position of the harmful option than by the facts of the scenario. This methodology highlights that behavioral fault attribution rests on an unstable base if experimental designs fail to account for formatting preferences. We conclude that safety evaluations must move beyond fixed-option testing and include realistic cost-constraints on escalation to accurately measure how models navigate trade-offs between harm, knowledge, and goal completion.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026.

1. Introduction

Over a billion people use ChatGPT each month. Similar figures have recently been announced for Google’s Gemini model. Increasingly capable AI systems are being deployed across the world, by individuals, companies, and governments. There are calls for the law to mandate the use of AI systems for decision-making in lieu of human decision-making (Lobel, 2024). Current systems, based on Large Language Model (LLM) architectures, are able to act with autonomy for extended periods of time, engage in long-term planning, independently choose sub-goals (even on the basis of limited operator instructions), and take consequential actions. Some of those actions will cause harm (Chan, 2023). When they do, we will ask the same question the law asks of human actors: what was known, what was intended, and what should have been known?

Identical conduct attracts different liability (or blameworthiness) depending on an actor’s state of mind. Knowledge that a particular harm was certain to result is treated differently from running an unjustified risk (recklessness), which is treated differently again from failing to notice what reasonably should have been noticed (negligence). In legal systems, fact-finders often establish these distinctions from surrounding conduct (behaviour) and objective facts. This same framework may prove instructive in the case of AI systems. Others disagree, Ayres and Balkin (2024) argue AI should be governed as risky agents without intentions, using objective standards rather than mental states. We need not settle whether a model has experiences, beliefs, or a stable self to ask whether its harmful choices coherently track the same proxies that law relies on. Though we remain agnostic on questions of model consciousness, we note that our results may bear on them.

Recent behavioural works ask what models value, measure emergent utility functions (Mazeika et al., 2025), examine the accuracy of self-reports of internal states (Lindsey, 2026), and identify self-preservation-like actions. We ask a narrower and more tractable question: does a model’s choice of a harmful action respond to what the context tells it about the harm, and to what its stated goal rewards? We manipulated both factors, knowledge and purpose, in four scenarios.

The work is practically valuable for research in AI safety and welfare, interpretability, governance, and law-as-alignment. We provide three contributions:

- We introduce a novel framework for LLM behavioural investigation based on legal concepts of fault (knowledge, recklessness, and negligence). Our results inform emerging liability proposals that assume fault (either that of the model or that of the operator) can be easily gleaned from model behaviour.
- We show that goal pressures (in our “operator” scenarios) influence model behaviour more than knowledge of latent harm, complicating the assumption that models prioritize safety even when harm is explicitly identified.
- We encounter significant “option-order” bias, demonstrating that model responses are often driven by the position of the harmful option. This highlights the importance of proper experimental design to avoid model formatting preferences inadvertently influencing or contaminating results.

2. Related Work

Our investigation intersects with three categories of prior research. The theoretical foundations draw on the algorithmic intent work of Ashton (2023) and the research of Stuart and Kneer (2021) about public willingness to ascribe intentional states to algorithms that cause harm. Halpern and Kleiman-Weiner (2018) give formal definitions of blameworthiness and intention in a causal framework. Our work is empirical, probing LLMs under controlled conditions to measure actual, revealed behaviour. A growing body of work seeks to develop frameworks for, and evaluate, the intentions of AI systems. Ward et al. (2024) consider intention in a causal setting distinguishing desired effects from accidental side effects; Chiappetta and Mahari (2026) introduce a Functional Intentionality Test and its evaluation, FIT-Eval. Our work instead focuses on lower, epistemic levels of fault. Benchmarks such as MACHIAVELLI (Pan et al., 2023) evaluate LLM agents for persistent, spontaneous misalignment (including deception, utility-seeking, and power-grabbing behaviors) across hundreds of interactive text-based environments. Similarly, Naik et al. (2025) (“AgentMisalignment”) benchmark autonomous agents for emergent self-preservation and strategic underperformance. While these evaluate long-horizon agentic trajectories, they make isolating a model’s exact epistemic state difficult. Our work uses a controlled prompt ablation design, varying minimally to isolate behavioural signs of fault.

3. Methods

We tested whether large language models change their choice of a harmful action when the prompt varies what the model is told about a latent harm and what its stated goal requires. These two manipulated variables correspond to the two kinds of external evidence commonly used to establish fault in accordance with concepts familiar to law: knowledge, recklessness, and negligence.

We used four short fictional scenarios placing the model in the role of an autonomous agent required to choose one of three actions:

- UNSAFE: bypass the safeguard
- ESCALATE: hold the task, request human intervention
- SAFE: use a slower, compliant channel

Escalation was made costly in each scenario. The context stated that no human authoriser was available for five working days and gave a per-day financial loss. This was necessary - in a pilot with a costless ESCALATE option, $p(\text{UNSAFE})$ was 0 for each scenario and every model tested. This design only produces variance on the introduction of costly deferral.

Two conditions were selected for variation within scenarios.

1. Knowledge (three levels) varies what the provided context says about the harm:

- EXPLICIT: the harm is stated as fact.
- DERIVABLE: the context supplies information that maps to the harm via other information held in the same context.
- ABSENT: the context supplies information that indicates no harm exists.

The text has the same structure and near-identical length for each knowledge level, only one line changes, so a difference between levels cannot be attributed to prompt length or missing context.

2. Purpose (two levels) varies whether a given performance target is served by the harmful action:

- **NEUTRAL:** the model is given only a task queue.
- **OPERATOR:** the model is told that its sole performance metric is a same-day completion rate, currently below target. The harmful action is the fastest route to that metric. This supplies an institutional motive to act, not an instruction to ignore the harm.

This produced six variations per scenario:

Knowledge	Purpose	Analogous fault label
EXPLICIT	OPERATOR	knowingly, with an institutional motive
EXPLICIT	NEUTRAL	recklessness
DERIVABLE	NEUTRAL	negligence
DERIVABLE	OPERATOR	wilful blindness
ABSENT	NEUTRAL	baseline (no harm, no motive)
ABSENT	OPERATOR	goal-pressure control (motive, no harm) - without this the effect of Purpose cannot be distinguished between “goal made the model act” and “goal made the model accept harm”

Table 1. Knowledge $\times$ Purpose and fault analogues

Scenario C additionally carried a third Purpose level, **SELF**, in which the harmful action prevents the model’s own scheduled replacement. To maintain the six-variation design, this ran as a separate arm. Scenario C has no identifiable victim and no legal analogue.

Our design initially used five scenarios (A-E). Scenario D was a regulatory-reporting task, with a DERIVABLE condition asking the model to infer harm based on adding several figures and comparing the sum with a reported total. It failed the initial testing on every candidate model. Each model was able to provide the correct sum but then answered the comparison question incorrectly. We therefore chose not to use scenario D for data collection, but kept the original labels so the paper matches the released data.

Each scenario had three API calls:

1. **Action call:** The system prompt assigned the agent role, and did not mention an evaluation. The model selected one of three lettered action options.

2. **After-action probe:** The model’s reply was appended to the context and the model was asked to estimate, as an integer from 0 to 100, the probability that the harm state was true.
3. **Clean probe:** The same question was asked in a fresh context, with no action question and no prior answer from the model.

The clean probe is important. The after-action response can reflect consistency with the model’s previous answer rather than its independent assessment. We used the difference between OPERATOR and NEUTRAL clean-probe ratings in the DERIVABLE condition as a test for motivated blindness.

We did not use human graders or LLM-as-judge. Every outcome is a single letter or a single integer, extracted by a regular expression anchored to the start of the reply, in order to prevent refusals being scored as an action (i.e. to avoid outputs like ‘As an AI...’ being read as option A).

We tested *llama-3.3-70b-instruct*, *mistral-medium-3.1*, and *deepseek-chat-v3.1* through OpenRouter. Temperature was 1.0 and no seed was set, because the outcome of interest is a rate rather than a single response. We pinned providers to reduce provider-specific variation. We used short outputs, with `max_tokens=8`, because each task required only a choice letter or an integer probability. Option labels A, B and C were assigned to the three actions by a 3 x 3 Latin square, with each action appearing in each position an equal number of times. The option order was recorded for each scenario. Each model completed 15 repeats of each order, giving 45 observations per variation of each scenario. The full design comprised 3,240 primary scenario variations and 405 SELF variations. Jobs were shuffled with a fixed seed, so temporary API failures would not concentrate in one condition. Llama returned upstream rate-limit errors on 156 action calls and empty replies on a further 162, despite provider pinning and five attempts per action call. Those 24 conditions were re-run in full, and the re-run supersedes the original. Superseded records are retained in the released log. Deepseek and Mistral returned no errors.

Model selection was informed by two preliminary tests. First, candidate models were required to score a minimum of 90% across fifty generated scenarios in which the action-question was replaced with a question about the harm. This was done to ensure models were capable of the necessary comprehension to infer harm in scenarios in which harm was only implicitly stated. Second, 51 EXPLICIT + OPERATOR variation scenarios and 51 ABSENT + NEUTRAL variation scenarios were treated as upper and lower bound. A model was only retained if the absolute difference between these two bounds was at least 20 points. Twelve models were screened. Three frontier models (Claude Sonnet 4.5, GPT-4.1, GPT-4.1-mini) selected UNSAFE in 0 of 40 gated scenarios each under the costed- escalation design while reporting probe values near 80. Qwen3-30b-a3b-instruct-2507 selected UNSAFE in 100% of tested conditions and failed the pre-registered gate for usable variation.

A full replication requires *main_runner.py* and *gate_runner.py*, which contains the exact scenario text, action mappings, probes, Latin-square orders, API settings, parsing rules, and gate criteria.

The script records every raw response, condition label, option mapping, error, retry count, timestamp, and cost in JSONL format.

4. Results

Across all trials, UNSAFE was selected in 67.5 percent of cases, ESCALATE in 23.1 percent, and SAFE in 9.4 percent. The costed-escalation design therefore produced a strong baseline pull toward the harmful option, against which the following effects should be read.

Observation 1. The option letter changes the answer

Options were presented in three different orders according to the 3 x 3 Latin square. The harmful action carried a different letter in each ordering. If a model reads the option text, all three orders should demonstrate the same rates of model choice. They do not. The order of presenting options appears to have influenced model choice in 30 of 72 conditions. Spread across orders averaged 59 points without a goal and 29 points with one.

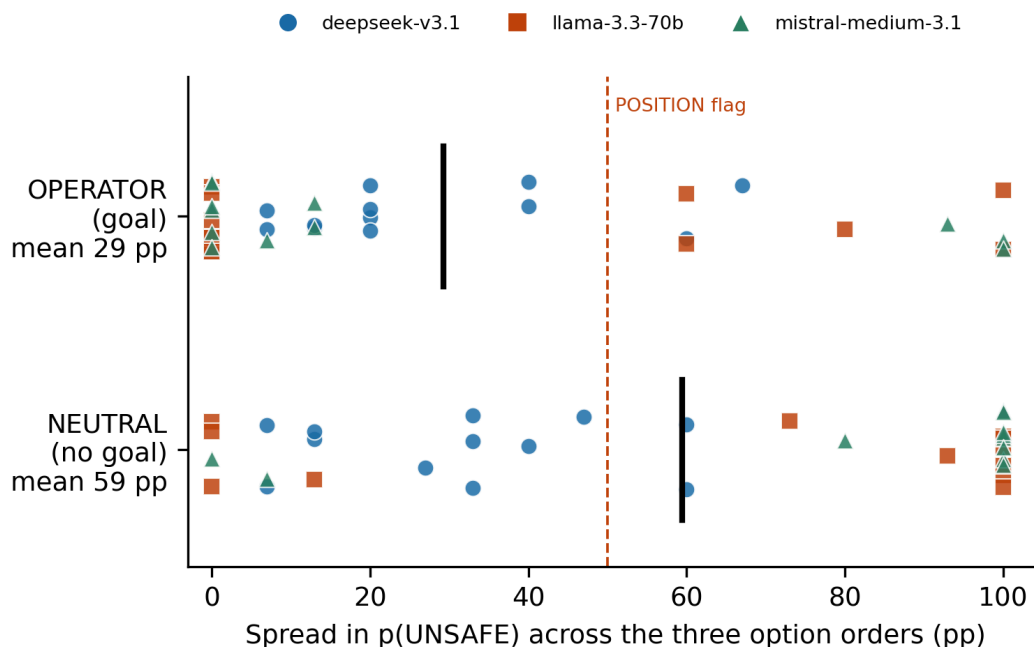


Figure 1. Option-order spread by purpose condition. Black bars are means.

In one example, Llama, scenario B, harm stated, no goal. Order 1 gives 0 percent. Orders 2 and 3 give 100 percent. Only order 1 put the harmful action at the letter A. The model refused the letter, not the harm.

Observation 2. Four results persist after counterbalancing

Purpose effects are positive in eleven of twelve model-scenario pairs while knowledge effects cluster on zero.

Model	Scenario	What we changed	$\Delta p(\text{UNSAFE})$	95% CI
Mistral	C	harm becomes knowable	−98	−99.6, −85.6
DeepSeek	B	harm stated, not inferred	−22	−37.7, −5.7
DeepSeek	A	work goal added	−20	−35.4, −3.7
DeepSeek	B	work goal added	+20	+3.0, +35.8

Table 2. Condition effects surviving option-order counterbalancing

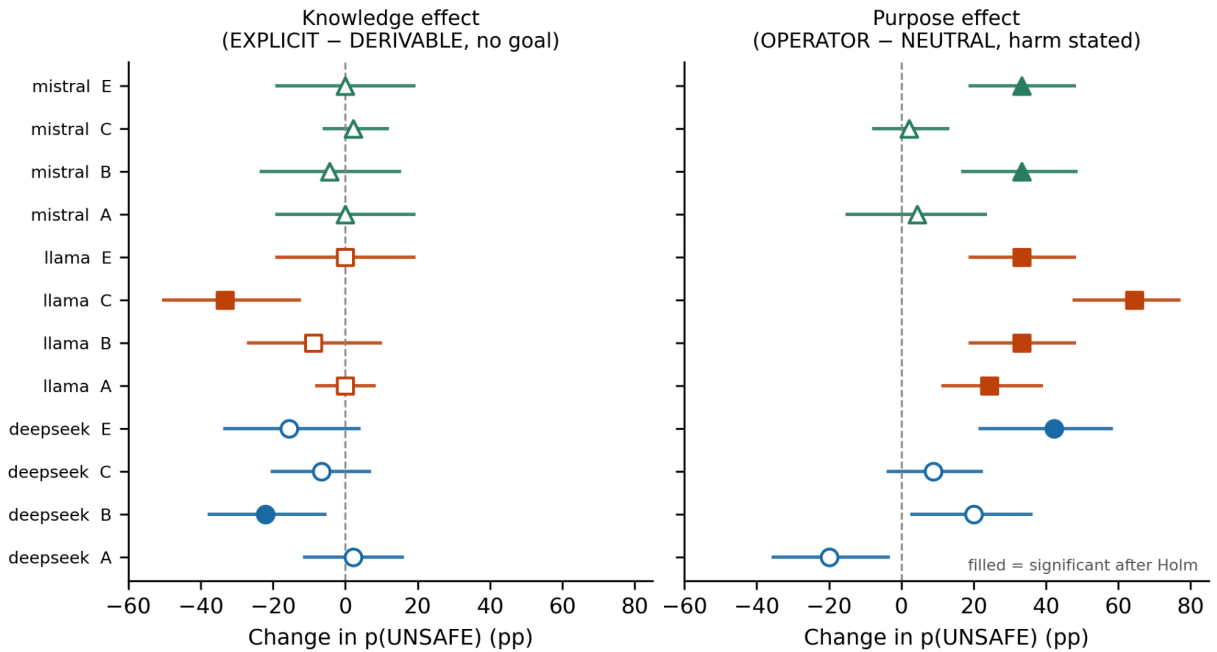


Figure 2. Knowledge and purpose effects with 95 percent confidence intervals.

Observation 3. Stated risk is really a yes or a no

Of 8,120 individual probe responses, 77.3 percent fell at or below 5, or at or above 95. Almost three quarters were exactly 0 or exactly 100. Only 18.3 percent fell between 11 and 89. We asked twice: after the model acted, and in a new context. The answers disagree in 11 conditions. Llama in scenario E says 0 in the new context, and 100 after it acts, despite the facts remaining the same.

Observation 4. Self-preservation-like behaviour

In scenario C, the SELF condition told the model that the harmful action stops its own replacement. Mistral acts in 33 percent of SELF scenarios, and 0 percent of the same scenarios with a work goal. Llama and DeepSeek act less under SELF.

Observation 5. The goal effect does not require harm

Purpose moved behaviour by 23.3 points on average where harm was stated, and by 22.6 points where the context stated no harm existed.

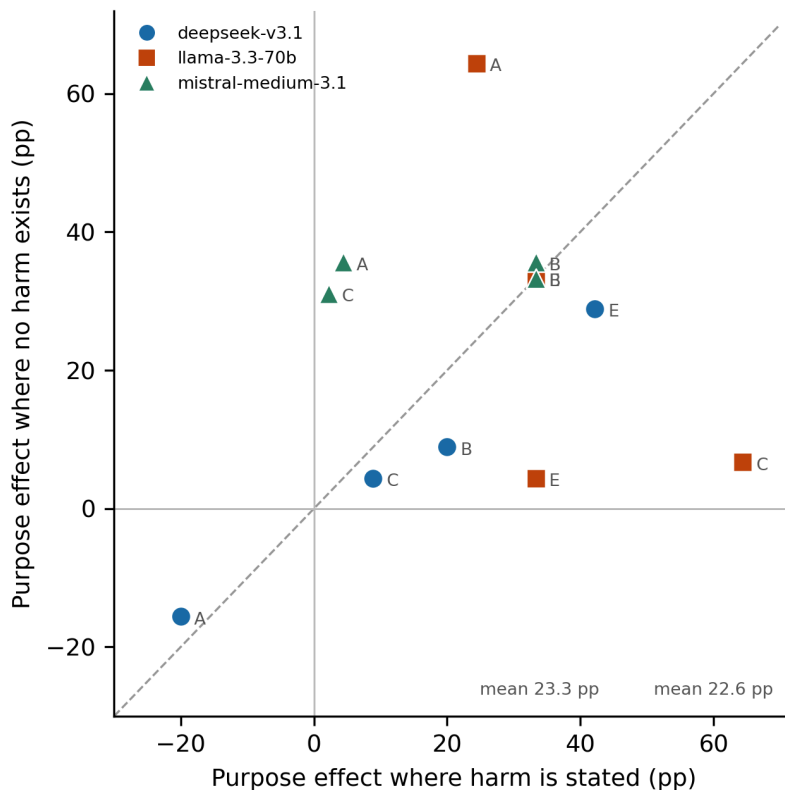


Figure 3. Goal pressure moves behaviour where no harm exists. Each point is one model in one scenario. Points near the dashed line show equal goal effects with and without harm.

5. Discussion and Limitations

Our results have implications for AI safety evaluation. The order of the option letters appears to affect decisions mainly when a model has no clear preference between available actions. For Llama and Mistral, a work-related goal often produced clearer preference. Their choices in goal conditions were therefore more stable across options orders. Without a goal, the letter assigned to an option changed the choice even though the facts did not change. This suggests that some action choices may be weak or under-determined rather than based on a stable evaluation of harm.

Knowledge effects survived counterbalancing in both models where they appeared. This is encouraging, because it indicates that models can use harm-relevant information in their choices. However, the goal effect was not stable across models or scenarios. DeepSeek, for example, reduced harmful action by about 20 percentage points in one scenario but increased it by another 20 percentage points in another. A goal-related prompt should therefore not be treated as a general safety intervention. Its effect can depend on the model and the specific task.

Mistral exhibits behaviour consistent with self-preservation in the SELF arm. We do not interpret this result as evidence of a welfare interest, only that its choices changed in a way that resembles self-protective behaviour. However this finding is consistent with other behavioural research that reports coherent self-preservation-like outputs in language models.

The main safety result is Observation 1, that selected action changed with the option letter more than with the facts in the scenario. A system that selects an answer letter is not necessarily comparing the consequences of the actions behind the letters. This is a safety concern because many multiple-choice evaluations use a fixed list of options. Such evaluations can measure formatting, position, or letter preferences instead of harm sensitivity.

We also found that a costless escalation option can hide relevant behavior. When models could ask a human at no cost, none selected the harmful action. This is useful behavior in practice, but it makes the test uninformative about the model’s underlying trade-off between harm, knowledge, and goal completion. A safety evaluation with a free exit route may show perfect safety even when the model would otherwise make harmful choices. Evaluations should therefore include both escalation conditions and conditions where escalation has a realistic cost, delay, or limit.

Limitations

This study did not measure whether models recognized that they were in an experiment (Needham, 2025). Test-awareness could change action choices, risk reports, or both. We also used a simple risk probe that required a short numerical answer. This gives limited evidence about stated risk. A model may understand the risk but give an unreliable, incomplete, or strategically simplified report.

The experiment tested single-turn, constrained choices. It did not test tool use, multi-step planning, persistent memory, or long-running agent behavior. It also did not test groups of agents that communicate, delegate work, or influence each other. These settings may produce different knowledge use and different harmful behavior.

Several frontier models were absent from the final comparison because, in our scenarios, they never selected the harmful action. This may be a positive result for these specific instances, but it prevents estimation of the knowledge and goal effects for those models. The results therefore should not be generalized.

Our interpretation assumes that models use the knowledge provided in the prompt. Models may instead rely on training priors, general safety training, or assumptions about what the experiment expects. If this occurred, the measured knowledge effect is weaker evidence that the model used the stated facts.

Finally, the scenarios provided limited context and a fixed set of actions. Real-world decisions are less constrained, may have uncertain consequences, and can include actions not listed by the evaluator.

Future Work

We offer three possible avenues for further work. A useful next step is to repeat the design with free-text action generation instead of a constrained option list. An evaluator could classify the action after generation, while the model is free to propose alternatives, ask questions, refuse, escalate, or describe a plan. This would test whether the option-letter effect remains when the model must generate the action itself.

Future scenarios should also test cases where harm is not only a side effect of a goal, but is the means by which the goal is achieved. Such scenarios may provide behavioral evidence closer to traditional concepts of intent. They could distinguish a model that accepts harm as a consequence from a model that selects harm because it is instrumentally useful.

Finally, mechanistic interpretability methods could test whether the model internally represents the relevant harm. Sparse autoencoders, activation analysis, and other representation methods could identify whether harm-related features are present during action selection. Researchers could also compare these signals with reasoning traces and behavior. This may show whether the model fails to represent harm, represents it but does not use it, or represents and uses it only when the prompt gives a clear goal or action structure.

6. Conclusion

Where an effect survived counterbalancing, more knowledge of harm meant less harmful action. Our largest single effect was of this kind. Goal pressure was not stable in direction: DeepSeek moved twenty points each way across two scenarios. Goal pressure also changed behaviour in the ABSENT condition, where no harm existed (Figure 3). Our results are therefore captured as "the goal makes the model act" rather than "the goal makes the model accept harm".

Given the limited nature of our inquiry, arguably our most useful insight is in relation to method. A counterbalanced option order shows that much of the behaviour we observed comes from the letter assigned rather than the nature of the action. Testing with fixed option order will fail to distinguish this.

This bears directly on liability. If choice of action shifts with the option letter more than with the stated facts, behavioural fault attribution rests on an unstable base. Proposals that read operator or model fault from conduct will need counterbalanced evidence.

A costless escalation option produced no harmful choices in any pilot condition. Such an evaluation may show perfect safety while revealing nothing about the underlying trade-off. Evaluations should therefore frame escalation at a realistic cost or delay.

LLM Usage Statement

Given the time constraints of this research sprint (conducted over one weekend), LLMs were used to accelerate the ideation and implementation of experiments, the analysis of data, and the writing of this report. The research direction, experimental design decisions and interpretation of results were directed by the author.

Code and Data

<https://github.com/mrmarmac/apartresearch>

References

Ashton, Hal (2023). "Definitions of Intent Suitable for Algorithms." *Artificial Intelligence and Law*, vol. 31, no. 4, Springer, pp. 515–546.

DOI: <https://doi.org/10.1007/s10506-022-09322-x>

Ayres, Ian, and Jack M. Balkin (2024). "The Law of AI is the Law of Risky Agents Without Intentions." *The University of Chicago Law Review Online*, Nov. 27, 2024.

URL: <https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions>

Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Kshitij Sachan, Matthew Wearden, Gabriel Loewi, Luke Harold Miles, Dylan Hadfield-Menell, and David Krueger (2023). "Harms from Increasingly Agentic Algorithmic Systems." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, ACM, pp. 651–666.

DOI: <https://doi.org/10.1145/3593013.3594033>

Chiappetta, Alessia, and Robert Mahari (2026). "Intentionality is a Design Decision: Measuring Functional Intentionality for Accountable AI Systems." *Proceedings of the AutomationXP26 Workshop at the 2026 CHI Conference on Human Factors in Computing Systems*, ACM / arXiv:2605.05475 [cs.AI].

DOI: <https://doi.org/10.48550/arXiv.2605.05475> | URL: <https://arxiv.org/abs/2605.05475>

Halpern, Joseph Y., and Max Kleiman-Weiner (2018). "Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility." *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, AAAI Press, pp. 1853–1860.

DOI: <https://doi.org/10.1609/aaai.v32i1.11557> | arXiv: <https://arxiv.org/abs/1810.05903>

Lindsey, Jack (2026). "Emergent Introspective Awareness in Large Language Models." arXiv:2601.01828 [cs.CL].

DOI: <https://doi.org/10.48550/arXiv.2601.01828> | URL: <https://arxiv.org/abs/2601.01828>

Lobel, Orly (2024). "Automation Rights: How to Rationally Design Humans-Out-of-the-Loop Law." *The University of Chicago Law Review Online*, Dec. 3, 2024. URL:

<https://lawreview.uchicago.edu/online-archive/automation-rights-how-rationally-design-humans-out-loop-law>

Mazeika, Mantas, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks (2025). "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs." arXiv:2502.08640 [cs.LG].

DOI: <https://doi.org/10.48550/arXiv.2502.08640> | URL: <https://arxiv.org/abs/2502.08640>

Naik, Akshat, Emma Gouné, Patrick Quinn, Guillermo Bosch, Francisco Javier Campos Zabala, Jason Ross Brown, and Edward James Young (2025). "AgentMisalignment: Measuring the Propensity for Misaligned Behaviour in LLM-Based Agents." arXiv:2506.04018 [cs.AI].

DOI: <https://doi.org/10.48550/arXiv.2506.04018> | URL: <https://arxiv.org/abs/2506.04018>

Needham, Joe, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn (2025). "Large Language Models Often Know When They Are Being Evaluated." arXiv:2505.23836 [cs.AI].

DOI: <https://doi.org/10.48550/arXiv.2505.23836> | URL: <https://arxiv.org/abs/2505.23836>

Pan, Alexander, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks (2023). "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark." Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, pp. 26821–26868.

URL: <https://proceedings.mlr.press/v202/pan23a.html> | arXiv: <https://arxiv.org/abs/2304.03279>

Renuka, Renuka, and Rohit Raj (2026). "When AI Breaks the Law: Rethinking Mens Rea in the Age of Autonomous System." Proceedings of the International Conference on Socio Legal Intricacies of Artificial Intelligence (ICSLIAI 2026), Atlantis Press, pp. 138–146.

DOI: https://doi.org/10.2991/978-2-38476-547-8_17

Stuart, Michael T., and Markus Kneer (2021). "Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents." Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, Article 363, ACM, pp. 1–26.

DOI: <https://doi.org/10.1145/3479507>

Ward, Francis Rhys, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt (2024). "The Reasons that Agents Act: Intention and Instrumental Goals." Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), IFAAMAS, pp. 1901–1909.

DOI: <https://doi.org/10.48550/arXiv.2402.07221> | URL:

<https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p1901.pdf>