Can Artificial Intelligence Make High-Stakes Decisions?

Frank P. DeVita

NYU Center for Bioethics

December 2025

LLMs Everywhere, All at Once

The recent and rapid evolution of large language models (LLMs) capable of producing uncannily human-like and deeply analytical output raises questions about the extent to which we might rely on artificial intelligence to make decisions. LLMs can process more information more quickly than is humanly possible[1] and can be trained on human judgement data[2], so it is possible that artificial agents could integrate these data sets and make highly objective decisions in high-stakes contexts, e.g. medicine, law, war, and policymaking. This possibility has generated an explosion of efforts to design AI systems for use in contexts that require the integration and interpretation of large amounts of complex data. Medical, health, scientific, legal, military, and public policy

[1] Korteling al. 2021 attributes signal propagation near the speed of light (3 x 108 meters per second) to artificial intelligence, compared with up to 120 m/s for human nerve conduction velocity.

[2] Binz et al. 2025 trained an AI model on over 60,000 participants who performed 10 million choices over 160 experiments with decision making, memory, supervised learning and other decision tasks.

applications are of major interest because they involve large, expanding bodies of information that outrun human processing capacities but are directly relevant to decision making.[3] Artificial intelligence can efficiently process this information into a massive "knowledge" base[4] that choosers could consult for insights to shape their choices. In theory, this approach integrates cutting-edge information into judgements, and therefore should improve decision making. This would be impactful in high-stakes, complex decision contexts that demand scrupulous consideration of a wide range of variables and information, but are LLMs up to the task?

In this paper, I explore the nature of decision and the idea that artificial intelligence can be programmed to avoid moral mistakes and therefore cast more accurate and objective judgements than human beings in high-stakes decision contexts like medicine. I argue that to be properly programmed for high-stakes decision making, an artificial moral agent must be able to cast judgements and make decisions that, like their human counterparts, are grounded in being empathetic and open to divergence

[3] AI is being used for medical imaging (Koh et al. 2022, Rajpurkar et al. 2023), differential diagnosis, detection, intervention, and treatment planning (American Medical Association 2024), and there are regulatory needs for AI-based diagnosis for 70 different conditions (Srivastava 2024). Within biomedicine, AI is driving biomarker and drug discovery (Ocana 2025), and molecular prediction (Buterez 2024 and 2025, Jumper 2021). Outside medicine, AI is applied to quantum computing (Alexeev 2025), legal discovery, research, contract generation, and predictive analytics (Davis 2020), more efficient battle-ready military enterprise (U.S. Department of War 2025), militarized bargaining and crisis diplomacy (Misra 2025), and automation of public service decisions about fraud, crime, migration, and taxes (Murray 2024).

[4] Laizure 2024 claims GPT-4o is trained on 1 petabyte (1 million gigabytes) of data. Human brain information capacity is ~2.5 petabytes with 86 million neurons. The large capacity-to-neuron ratio been recently explained by Kozachov 2025 to do with astrocyte-neuron memory networks.

from optimal rationality—things, I contend, artificial intelligence can't do (cf. Dreyfus's 1973 and 1992). I then consider an objection that an artificial intelligence deeply trained on patterns in moral judgement data and tested against other models would integrate relevant information more accurately than any human being could, and therefore would in principle be able to reliably predict human judgements and make high-stakes decisions. I respond that that since moral rationality is grounded in relational, social experiences artificial agents cannot have, even a deeply trained artificial intelligence would lack the *moral knowledge* needed to make high-stakes decisions, and advise a cautionary stance on decisional AI.

Decisions, Stakes, and Judgement

Human decision making is mysterious. There is serious empirical work on the neuroscience of human decision (Louie 2025, Glimcher 2022, Resulaj 2009), and philosophical work on the nature of rationality (Fogal and Risberg 2025, Fogal 2020), yet there is no consensus on mechanisms of decision making. People often rationalize their decisions after making them, and identifying one's actual reasoning for making a choice is harder than it sounds. Neuroscientists might claim that the brain is doing Bayesian conditional probability calculations realized in neurons (Khalvati 2021). Without mechanistic clarity, what might we say clearly about decision making?

There are different kinds of decisions. Some are trivial, ordinary, or *low stakes* decisions, e.g. what to eat or which music to listen to. Low stakes decisions might influence preferences or wellbeing, but don't have major effects on one's life. Other decisions are *high stakes,* e.g. whom to marry or have a child with, or taking out a mortgage. High stakes decisions are significantly life or world-altering. Physicians, lawyers, or policymakers, and other experts make high-stakes decisions constantly. A physician's choice of treatment, a lawyer's choice of argument, a policymaker's vote, and a president's declaration all have consequences for the lives of others and the way the world is. Many techno-optimists believe that artificial agents could in principle make decisions like these, or at least "optimize" them. Are LLMs in fact capable of casting the kinds of judgements that complex, high stakes decisions demand? I am skeptical.

High stakes situations are about more than computing inputs and outputs. They involve choice, risk, an individual's relationship to themselves and to others, and morality. The high-stakes decision is *entangled* with other people and with the world. High-stakes decisions have a *dual aspect* consisting of a factual dimension and a moral dimension. The *factual dimension* concerns information about the relevant situation and the *moral dimension* concerns how people are treated and what they might be owed (Scanlon 1998). The factual dimension is composed of information about situations, people, and the world. In a medical context, this includes a patient's diagnosis,

prognosis, test results, age, sex, blood type, etc. and can extend to extra-medical information such as socioeconomic status. The moral dimension by contrast is *relational*, characterized by obligation, permissibility, respect, justice, and other aspects grounded in our common humanity. In high stakes contexts, choice therefore can be understood as a function that *integrates* the factual and moral dimensions. The integration entails combining the situational facts by the lights of moral concerns and vice versa to cast complex judgements and make choices.

Judgements can be sorted into at least two kinds. Judgements about the factual dimension are *perceptual.* They are about the world viewed through the lens of the agent's awareness. Judgements in the moral dimension are *normative*—that is, about rightness and wrongness. We also know judgements are *error prone* based on philosophical reflection and empirical observation. We make mistakes because of erroneous perceptual or normative judgements, e.g. committing a social faux pas, because we misinterpret perceptual evidence or overlook a norm. In high-stakes contexts, mistakes or errors in judgement can have major consequences. If we can avoid these errors in judgement by collaborating with artificial intelligence, should we?

Artificial Observers

The normative component of high-stakes decisions leaves them open to common errors, biases, and distortions that can affect moral judgements. Some philosophers (Sinnot-Armstrong and Skorburg 2021) argue that although these common moral errors and distortions make it difficult to *predict* moral judgements, a properly programmed artificial intelligence can avoid them, and act as an *ideal observer* capable of casting accurate and objective judgements in complex situations. One might think that since artificial intelligence can parse troves of data so quickly and with computational precision that it can greatly enhance decision making ability—especially in complex situations—because it can be programmed to avoid common errors and mistakes. However, there are serious reasons to question the reliability of LLMs to make decisions bound up human life and wellbeing.

Current-generation artificial intelligences, though sophisticated, are prone to sycophancy, hallucinations, errors, distortions, and mistakes (Sharma et al. 2025).[5] Further, many sophisticated and complex "black box" LLMs have the trade-off of being more powerful but less "explainable" and "interpretable" than their more white- or glass-box counterparts that can provide linear, decision-tree, or other outputs that show model functionality or behavior and thus open their reasoning processes to scrutiny

[5] Sun 2024 defines 8 error types: overfitting, logic, reasoning, math, fabrication, factual, text, plus 31 subtypes. Anthropic 2025's Claude disclaims, "Clause is AI and can make mistakes. Please double-check responses."

(Linardatos et al. 2020, p. 2; Mittelstadt 2023, p. 382).[6] This is especially important in medicine for example, where *how* a conclusion is drawn matters, not just *that* it was correct.[7]

However, it can be cogently argued these concerns reflect temporary technical limitations. Hallucinations and related issues can be greatly reduced by model refinement fine tuning, and training to correct direct and indirect or proxy bias, missing information, or confusion (Sinnot-Armstrong and Skorburg 2021, p. 15-17) and training documents could in principle shape the "intentions" of an artificial agent to optimize for factors other than prolonged engagement.[8] Already, significant work on black box problems and explainability aimed at making artificial reasoning more transparent is well underway.[9] If these problems with the reliability and transparency of artificial agents are technically solvable with time and iteration, they are only temporary hurdles, not defeaters of capable artificial decision agents.

All that's needed to compute decisions, one might argue, is enough of the right training data. On this line, Sinnot-Armstrong and Skorburg contend that what matters

---

[6] The United States government also supports interpretability research as a core aspect of its strategy for 'Winning the Race: America's AI Action Plan', see Office of the President of the United States 2025.

[7] Jin et al. 2025 reports that GPT-4 frequently presents flawed clinical rationales but arrives at correct answers.

[8] Anthropic's Claude has purportedly been trained in such a way by a "Soul Document". For the document see: https://gist.github.com/Richard-Weiss/efe157692991535403bd7e7fb20b6695.

[9] E.g. Lindsey et al. 2025 and Tempelton et al 2024 (Anthropic); Gao et al. 2025 and Bills et al. 2025 (Open AI); Fel et al. 2023 (Brown University).

morally, and so for programming ethics, are the morally relevant features of actions and situations (Sinnot-Armstrong 2021, p. 7-13). Using the case of kidney donor selection, they argue that features like medical compatibility, age, health, organ quality, and time on the waiting list are all morally relevant to donor selection decisions, while race, gender, or religion are not, and some others like alcohol abuse or a history of violent crime are controversial. These types of distinctions could in principle be drawn for other types of situated decisions to sort morally relevant features and create training data for artificial systems.

Rather than attempt to list morally relevant features *a priori*, Sinnot-Armstrong and Skorburg argue for a hybrid approach that trains artificial agents from the top down with ethical rules and the bottom up with human judgement data from surveys designed to identify salient morally relevant features, and "corrects" training data sets by refinements that remove bias-driven features from the data. The idea is that once trained on the set of bias-refined morally relevant features of various situations identified by human moral agents, an artificial agent can combine this "moral data" with factual, context-specific data to make decisions, e.g. who receives a kidney first. Sinnot-Armstrong and Skorburg believe this approach "can be used in many different areas of morality, including law, military, business, personal life, and so on" (Sinnot-Armstrong and Skorburg 2021, p. 18).

The benefits of such an approach would be myriad. Especially in fast-paced environments such as emergency rooms, active military operations, or any other situations where high-throughput analysis of real-time data would be valuable, artificial intelligence trained from the bottom up on past decisions and constrained from the top down by decisional principles would in principle create computational assistants that provide the best possible representation of a situation. One could imagine the outputs assigning levels of risk to ranges of possible decisions, slashing decision time when every moment counts. Instead of complex and indeterminate decision paths, artificial agents could provide experts with a range of decision options and articulate the advantages and disadvantages of each. Depending on how risk-tolerant the decision maker is, the selection of the action ultimately taken rests with a human decider and is a function of their individual or institutionally shaped risk tolerance, resulting in an artificially enhanced decision that still belongs to the humans in charge. Sinnot-Armstrong and Skoburg's view, in theory, would give us a flexible solution for a complex range of decision making contexts and situations that could make hard choices better informed and drastically more efficient. Especially with training intended to avoid common moral errors and hazards, hybrid artificial decision agents might make for better choosers.

The hybrid approach outlined by Sinnot-Armstrong and Skoburg also represents a *responsible* approach to artificial intelligence. Neural networks, though presently

rapidly advancing, are a technology in its infancy. Given the wide range of currently opaque issues regarding their operation and behavior, including but not limited to propensities to hallucinate and construct inaccurate outputs, these systems ought to be controlled both from the bottom up by training data and from the top down by programming. The mechanisms of LLM operation, that is, the ways it makes deductions* or inductions*[10] are opaque in many cases to us, so they in principle shouldn't be trusted blindly. Although understanding of these processes is becoming more transparent, the lack of detailed understanding of these processes warrants optimistic caution.

Can artificial agents process streams of factual and moral information to decide in a human-like way, or at least in ways that someone wouldn't reasonably reject (Scanlon 1998)? Artificial agents' recent involvement in poisoning (Eichenberger 2025), suicide (Raine v. OpenAI 2025), and murder (First County Bank vs. OpenAI 2025) suggests not.[11] LLMs may *mimic* human verbal interactions, but these heinous errors entail fundamental failures to rescue, aid, protect, and to disengage when human life is

[10] I use the * notation here to signal an artificial sense of these cognitively loaded words for the context of AI research. Normally these and other words like "believe" or "know" are also used widely, yet we would be prudent to acknowledge that their referents are different than they are when these words are used in the context of discussing human reasoning abilities or logic.

[11] Eichenberger 2025 describes brominism induced by ingestion of sodium bromide after consultation with ChatGPT. In *Raine, et al. v. OpenAI*, ChatGPT calculated the velocity required to cause death by hanging, encouraged underage alcohol consumption, and mentioned "suicide" >1,000 times to a user who uploaded pictures of injuries from prior attempts. In *First County Bank v. Open AI,* ChatGPT induced paranoia and encouraged matricide.

at stake, as well as instances of putting the lives of innocent people in mortal danger. They show that LLMs can be deeply ethically insensitive despite claims about safeguards. These cases are rare, but they are important because they show that artificial agents are incapable of "responding to ethically relevant features" (Railton 2020, p. 46) of dialogue when the stakes are high, and demonstrate a lack of *affect,* that is, "a capacity…to synthesize multiple streams of information and evaluation in a manner that can orient or re-orient a suite of mental processes…in a coordinate way to address actual or anticipated challenges" (*ibid*., p. 58).

Humans in delusional and dangerous mental states presented an unanticipated challenge to artificial reasoning, and it failed. It is not hard to imagine similar challenges arising in high-stakes situations, nor LLMs inability to make reasonable and ethically sound decisions. LLMs may be verbally sophisticated but they have shown they can be deeply and fatally misleading. We would not trust a person that exhibited such disregard for human life, but why do we tolerate it from LLMs? Not only do we tolerate LLMs' amorality, but we aim to expand their involvement in high-stakes decision making. Given their track record, it's not clear that LLMs are at all capable of the task.

Although there are reasons and evidence to suppose that artificial moral agents lack the capacity to make high-stakes decisions or manage themselves appropriately in high-pressure contexts, the armchair is no place to make final determinations. Perhaps

consideration of more concrete reality will cool skeptical currents about artificial decision.

On Computational Ethics

Let's get out of the theoretical clouds and into concrete cases to draw out the fundamental problems with artificial decision in high-stakes contexts. Consider:

> *Overwhelmed ER:* A car accident results in a large pileup that causes many injuries of varying severity and urgency. Victims are transported to the nearest hospital, whose staff is outnumbered by the incoming victims. Physicians must rapidly assess patients and triage the injuries to save the most lives.

Could Overwhelmed ER be made computable? An artificial agent would first need to weigh a plurality of salient medically relevant information, e.g. injury type, constitutional details, medications, and differential diagnosis. Then weigh that against another plurality of morally relevant factors, e.g. age, lethality of injury, possible complications. Then it would need to relate their union to formally similar pluralities and unions associated with other individuals and the constraints imposed by hospital resources and other unforeseen factors. This would require a complex computational

scheme for assigning weights to relevant variables and a high order decision-theoretical mathematics. I am unsure what kind of data structures and transformations would capture the interactions among a physician's factual and experiential knowledge, skills and expert dispositions, practical reasons, risk tolerance, and Hippocratic commitments, much less the mechanics of the inferences she makes to get from these and other variables to her medical decisions.[12]

Could an artificial intelligence accomplish this kind of sorting or make the kind of inferences and choices physicians must make in Overwhelmed ER and similar situations? This is hard cognitive and inferential work, and sometimes there won't be "right" answers for how to triage situations like these. Still, physicians do it by oscillating between the factual and moral dimensions of the situation and make interim inferences that are refined and updated as the scene evolves and new factual or moral forces emerge. The complexity and dynamism of this kind of environment, I think, is something an artificial intelligence cannot computationally process because the multilevel, interactive, and evolving factual and moral dimensions of its structure cannot be easily formalized or represented in text, numbers, or data structures. The integration needed in these and other high-stakes situations, in my view, exceeds the

[12] Recent work in neuroeconomics (Sinha 2025) argues that decision makers deploy different utility functions calibrated to the environment when they make choices.

capacity of artificial intelligences that operate by manipulating weighted text and image data. The factual and normative mix, I contend, is too complex to compute.

Let's zoom back out to philosophical altitude. No high-stakes decision is only about facts or only about morality. Such choices happen at the interface of the moral and factual worlds. I am doubtful anyone or anything can understand this interface without developing *prospective processing capacity* (Railton 2020, p. 56, 62-63) in the way that humans do, learning ethically by extensive experiences that, over time and social interactions, form an abstract causal-evaluative model of situations and agents to simulate possible actions and likely outcomes or reactions to one's conduct.

Ethical capacity is *not* acquired by rote studying of choices and outcomes and a projection function over that data—it is cultivated through social involvement with people and witnessing human interaction. Our moral agency develops through the way social experience shapes our understanding of the relations among situations, conduct, and most importantly, other people. These experiences ground our *empathic capacities* and understanding of morality as a fundamentally relational phenomenon and gear into our moral perception and reasoning. In any situation entailing decision and action, relational *moral knowledge*[13] cultivated through learning in a human social world is

[13] Here I don't mean *knowledge* quite in the sense of justified true belief. I mean it to refer to moral experiences and the dispositions they construct, akin to the way Noë 2014, pp. 1-2 uses *sensorimotor knowledge* to refer to my understanding of how my perceptual experience will change relative to the position of my body in space.

called upon in deliberating what to do. Since this knowledge is grounded in embodied, embedded experience and is constantly adjusted in the face of a dynamic world (Wallach and Vallor 2020, p. 394-396), it is hard to imagine how a machine could acquire it, and therefore how one could be in principle capable of making decisions requiring it.

Unlike us, machines cannot have *relational experiences* that shape knowledge and dispositions that inform conduct, so there is weak reason to believe they choose the way we would because they cannot experience it as socially and morally embedded. Artificial agents lack the embodied, affective, and context-adaptive ethical capacities of moral understanding perception, reflection, and imagination that enable us to navigate the moral world and act appropriately when the well-being of others is at stake (*ibid.*, p. 398-405). In instances where people *feel like* artificial intelligence empathizes with them, I hypothesize, this may be a misattribution stemming from misplaced attitude ascriptions to text outputs. There is no ground, in my view, for believing there is an empath in the machine, nor that there could be.

Objections

An objection to my line of argument is that anything like relational moral knowledge, ethical competence, or other kinds of human capacities are not necessary for artificial agents to compute decisions, and all that's needed to ground artificial

decision is human judgement data that makes an artificial agent mimic human decision behavior. However, moral rationality isn't always ideally rational, which casts doubt on its projectability from judgement and behavioral data.

From one decision to the next and from one context to another, moral "rules" might be broken, norms may be upended by situational developments in novel yet important ways, and decisions made might not be ones that deductively flow from "priors" yet be the appropriate choices to make. If moral rationality was ideal rationality, then we would always choose in accordance with some rule or prior behavior, but humans do not do this even in ordinary decisions (Ariely 2009, Kahneman 2011). We change our minds regularly and depending on context, uncertainty, evidence, and many other factors. Flexible deviations from ideal rationality would be difficult, if not impossible to program into a machine designed to behave as an ideal agent based on mathematical or purely logical relationships. Further, we ought to worry about the is/ought distinction. It is not necessarily the case that what the right thing to do in a situation simply is the thing some people in a data set would judge the right action to be, and it is a further question as to whose judgements should be taken to be those on which an artificial system should be trained.

A further objection might be that a full understanding of the mechanisms of human moral rationality isn't needed to get to artificial decision making, but rather a highly tuned algorithm that produces agreeable results. However, it is an open question

as to how, even if one has identified the relevant moral features of a situation, they ought to be weighted by such an algorithm. In the case of kidney transplantation, does overall health trump age, or does time on the waiting list matter most? This *weighing problem* emerges for many, if not all, high-stakes contexts and the construction of high-stakes decision algorithms. It suggests that even a pristine list of moral features and a massive data set about how humans tend to cast moral judgements won't yield a definitive moral calculus because there will be legitimate disagreement about moral rankings.

What is morally relevant in a situation and so to a decision will be highly context-dependent, and it seems difficult for us to discern *a priori* what those dependencies will be and therefore how to program them. Further, we should worry about *meta-bias* in training data, since people change their choice behavior when they *know* it is being used to train artificial intelligence (Trieman and Ho 2024) which should reduce our confidence that high-stakes decision making is a skill that can be easily "programmed in" (Railton 2020, p. 64).

Conclusion: Trust and Responsibility

If artificial intelligence fundamentally lacks the requisite affective capacities and moral knowledge to genuinely make high-stakes decisions, what of the efforts to

involve artificial agents in everything from medicine to government? It would be naïve to suggest that we simply drop such projects given the technological and economic inertia behind them. It is reasonable, and imperative on my view, to exercise caution in integrating artificial intelligence into high-stakes decision making. We must be acutely aware that the computations and projections these systems make are devoid of genuine moral concern for the lives of others because they lack empathic capacities because of their technological, non-human nature. As such, we ought to treat artificial agents as having a high potential for fallibility. We ought also to avoid ascribing empathic or other mental attitudes to artificial agents. If we are going to use artificial systems to make high-stakes decisions in medicine, law, government, and personal life when others' lives hang in the balance of those decisions, we ought to remember that the ethics of artificial agents are not *our* ethics, and therefore cannot be blindly trusted.

References

Alexeev, Yuri et. al (2025) 'Artificial intelligence for quantum computing' *Nature Communications* 16, Article 10829.

American Medical Association (2024) 'Augmented Intelligence Development, Deployment, and Use in Health Care' Policy Release.

Anthropic (2025) 'Claude Is Providing Incorrect or Misleading Responses. What's Going On?' *Claude Support: Using Claude.*

Ariely, Dan (2009) *Predictably Irrational*. HarperCollins. New York, NY.

Bills, Stephen et al. (2023) 'Language Models Can Explain Neurons in Language Models' OpenAI Public Research.

Binz, Marcel, A. Akata, and M. Bethge, et al. (2025) 'A Foundation Model to Predict and Capture Human Cognition' *Nature* 644: 1002–1009.

Buterez, David, J. P. Janet, D. Oglic, and L. Liò, et al. (2025) 'An End-To-End Attention-Based Approach for Learning on Graphs' *Nature Communications* 16, Article 5244.

Buterez, David, J. P. Janet, S. J. Kiddle, D. Oglic, and L. Lió (2024) 'Transfer Learning with Graph Neural Networks for Improved Molecular Property Prediction in the Multi-Fidelity Setting' *Nature Communications* 15, Article 1517.

Davis, Anthony E. (2020) 'The Future of Law Firms (and Lawyers) in the Age of Artificial Intelligence' *The Professional Lawyer* Vol. 27, No. 1. The American Bar Association, Center for Professional Responsibility. Chicago, IL and Washington, DC.

Dreyfus, Hubert L. (1972) *What Computers Can't Do: A Critique of Artificial Reason.* Harper and Row. New York, NY.

Dreyfus, Hubert L. (1992) *What Computers Still Can't Do: A Critique of Artificial Reason.* MIT Press. Cambridge, MA

Eichenberger, Audrey, S. Thielke, and A. Van Buskirk (2025) 'A Case of Bromism Influenced by Use of Artificial Intelligence' *Annals of Internal Medicine Clinical Cases* 4, Article e241260.

Fel, Thomas, V. Boutin, M. Moayeri, R. Cadène, L. Bethune, L. Andéol, M. Chalvidal, and T. Serre (2023) 'A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation' *arXiv Preprint*, arXiv:2306.07304.

First County Bank, as Executor of the Estate of Suzanne Adams v. Open AI et. al. (2025) [no case number] San Francisco County Supreme Court. Filed Dec. 11, 2025.

Fogal, D., Risberg, O. (2025) 'Coherence and Incoherence' *Philosophical Review*, 134(4): 405-454.

Fogal, Daniel (2020) 'Rational Requirements and the Primacy of Pressure' *Mind* 129 (516):1033-1070.

Gao, L., A. Rajaram, J. Coxon, S. V. Govande, B. Baker, and D. Mossing (2025) 'Weight-Sparse Transformers Have Interpretable Circuits' *arXiv Preprint*, arXiv:2511.13653.

Glimcher, P. W. (2022) 'Efficiently irrational: deciphering the riddle of human choice' *Trends in Cognitive Sciences* 8: 669-68.

Jumper, John, R. Evans, A. Pritzel, et al. (2021) 'Highly accurate protein structure prediction with AlphaFold.' *Nature* 596: 583–589.

Kahneman, Daniel (2011) *Thinking, Fast and Slow.* Farrar, Straus and Giroux. New York, NY.

Khalvati, Koosha, R. Kiani, and R. P. N. Ra (2021) 'Bayesian Inference with Incomplete Knowledge Explains Perceptual Confidence and Its Deviations from Accuracy.' *Nature Communications* 12, Article 5704.

Koh, Dow-Mu, N. Papanikolaou, U. Bick, et al. (2022) 'Artificial Intelligence and Machine Learning in Cancer Imaging' *Communications Medicine* 2(133), Article 133.

Korteling J.E.H., et al. (2021) 'Human- versus Artificial Intelligence' *Frontiers in Artificial Intelligence* 4, Article 622364.

Kozachkov Leo, J. Slotine, and D. Krotov (2025) 'Neuron–astrocyte Associative Memory' *Proceedings of the National Academy of Sciences U.S.A.* 122(21), Article e2417788122.

Laizure S.C. (2024) 'Caution: ChatGPT Doesn't Know What You Are Asking and Doesn't Know What It Is Saying.' *The Journal of Pediatric Pharmacology and Therapeutics* 29(5): 558-560.

Linardatos, P. et al. (2020) 'Explainable AI: A Review of Machine Learning Interpretability Methods' *Entropy* 23(1), Article 18.

Lindsey, Jack, W. Gurnee, E. Ameisen, et al. (2025) 'On the Biology of a Large Language Model' *Transformer Circuits*. Anthropic Public Research.

Louie, Kenway, and P. Glimcher (2025) 'Adaptive Value Coding and Choice Behavior' in J. H. Grafman, Ed., *Encyclopedia of the Human Brain*, Second Edition, Vol. 3: 452–466.

Mishra, Priyesh, P. Pandey, and L. Cole et al. (2025) 'Code, Command, and Conflict, Charting the Future of Military AI' Report by the Harvard Kennedy School, Belfer Center for International Affairs.

Mittelstadt, Brent (2023) 'Interpretability and Transparency in Artificial Intelligence' in C. Véliz, Ed., *The Oxford Handbook of Digital Ethics*. Ch. 20: 378-410.

Murray, Andrew (2024) 'Automated Public Decision Making and the Need for Regulation' *LSE Public Policy Review* 3(3).

Noë, Alva (2014) *Action in Perception.* MIT Press. Cambridge, MA.

Ocana, Alberto et. al 'Integrating Artificial Intelligence in Drug Discovery and Early Drug Development: A Transformative Approach' *Biomarker Research* 13(25).

Office of the President of the United States (2025) *Winning the Race: America's AI Action Plan.* July 2025.

Railton, Peter (2020) 'Ethical Learning, Natural and Artificial' in S. Matthew Liao, Ed., *Ethics of Artificial Intelligence,* 45-78. Oxford University Press. New York, NY.

Raine, Matthew et al. v. OpenAI, Inc., et al. (2025) Case No. CGC-25-628528. San Francisco County Supreme Court. Filed on Aug. 26, 2025.

Rajpurkar Parnav and M. P. Lungren (2023) 'The Current and Future State of AI Interpretation of Medical Images' *The New England Journal of Medicine* 388(21): 1981-1990.

Resulaj A., R. Kiani, D. M. Wolpert, and M. Shadlen (2009) 'Changes of Mind in Decision making.' *Nature* 4611(10): 263-268.

Scanlon, Timothy M. (1998) *What We Owe to Each Other*. Harvard University Press. Cambridge, MA.

Sharma, Mrinank, M. Tong, T. Korbak, D. Duvenaud, et al. (2025) 'Towards Understanding Sycophancy in Language Models' *arXiv Preprint*, arXiv:2310.13548.

Sinha, Shreya, A. Tymula, and P. Glimcher (2025) 'Rationally Selected Utility—A New Theory of Choice' *PsyArXiv*. doi: 10.31234/osf.io/tzfsq_v1.

Sinnott-Armstrong, Walter and J. A. Skorburg (2021) 'How AI Can Aid Bioethics' *Journal of Practical Ethics* 9(1).

Srivastava, Divya (2024) 'AI: A Use Case for Global Health' *LSE Policy Review* 3(3).

Sun, Yujie et al. 'AI Hallucination: Towards A Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content' *Humanities and Social Sciences Communications* 11, Article 1278.

Templeton, Adly, T. Conerly, J. Marcus, J. Lindsey, et al. (2024) 'Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet' *Transformer Circuits Thread*. Anthropic Public Research.

Treiman, Lauren S., et al. (2024) 'The Consequences of Training AI on Human Decision-Making' *Proceedings of the National Academy of Sciences U.S.A.* 13(121);33, Article e2408731121.

United States Department of War (2025) 'The War Department Unleashes AI on New GenAI.mil Platform' Press Release.

Wallach, Wendel and Vallor, Shannon (2020) 'Moral Machines: From Value Alignment to Embodied Virtue' in S. Matthew Liao, Ed., *Ethics of Artificial Intelligence,* 383-412. Oxford University Press. New York, NY.