Vamshi Jandhyala

Encoding judgement

June 2026

Agents rarely lack judgement because the model is weak. They lack it because the judgement was never written down anywhere the agent can reach.

"AI cannot replace human judgement" is among the most repeated sentences in enterprise AI and one of the least useful. It is true in the way that "software cannot replace management" is true, and about as actionable. The sentence that earns its keep is the one underneath it: judgement has to be encoded before a machine can use any of it, and most of the time an agent fails not because the model lacked judgement but because nobody wrote the judgement down anywhere the agent could reach.

Take a credit analyst who glances at a borrower's file and says the numbers do not feel right. Press them and a reason will come, but it arrived after the verdict, not before it. The judgement was real and it was good, and it lived nowhere except in that analyst's head. An agent set to do the same screening starts with none of it. Whatever the analyst knew that was never written down, the agent cannot use, because for the agent there is no difference between knowledge that does not exist and knowledge that was simply never encoded.

A large model does arrive already steeped in judgement, having read more expertise than any analyst could. But what it read was the general literature, not your firm. The judgement that decides whether a screening is any good is the proprietary, local kind: which of your tables is point-in-time, what this desk treats as a red flag, where the data is known to lie. That judgement was never in the training corpus, and it is precisely the part still sitting unwritten in someone's head.

Judgement is the part that resists being written down

This is an old observation, older than computing. Michael Polanyi opened a 1966 book on it with a line that has outlived almost everything around it: we can know more than we can tell.[1] We pick a face out of a crowd and cannot say how we did it. The Dreyfus brothers narrowed the point to expertise: the novice follows rules, the expert has stopped following them and reads the situation directly, and force an expert

[1] Michael Polanyi, *The Tacit Dimension* (University of Chicago Press, 1966). The opening formulation is "we can know more than we can tell," and the recognition of a face we cannot describe is Polanyi's own example. Later writers call the idea Polanyi's paradox.

back onto explicit rules and their performance falls to the level of a careful beginner.[2] Judgement, in this sense, is exactly the part of knowing that resists articulation. Which is the whole difficulty, because articulation is the only way to hand it to a machine.

The first serious attempt to encode judgement walked straight into that wall. Expert systems in the 1980s set out to capture what a specialist knew as a base of explicit rules. A knowledge engineer would sit with the expert and try to draw the rules out, and the recurring obstacle was given a name that has aged well: the knowledge-acquisition bottleneck.[3] Feigenbaum named it in 1977, and the systems he had in mind he called applications-oriented intelligent agents. Two things kept breaking. The experts could not fully say what they knew, so the rules were always a lossy transcript of the judgement. And the rules that did get written were brittle, blind to the case nobody anticipated and slowly falsified as the world moved underneath them. The judgement decayed because it had been frozen.

The medium of encoding changed

What large language models changed is not that judgement became easy to encode. What changed is the medium. You no longer have to state the rule; you can supply examples and preferences and let the model infer the rule it could never be told. Reinforcement learning from human feedback, stripped of its machinery, is a way of encoding judgement by comparison: a person says this answer is better than that one, many thousands of times, and a reward model is fitted to the preference.[4] Constitutional AI takes a further step and writes the judgement down as an explicit set of principles, a constitution the model critiques itself against, trading a little of the implicitness back for something a person can actually read.[5] For the first time the old bottleneck loosened. You could encode a judgement you could not articulate, by showing instead of telling.

The bottleneck loosened, but it did not vanish. It moved. When judgement is encoded by example, you no longer know precisely which judgement you encoded, because it sits latent in the data rather than stated in a rule. The failure modes are the proof of it. A reward model trained on human comparisons learns whatever those

[2] Hubert L. Dreyfus and Stuart E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (Free Press, 1986). Their five-stage model holds that analytic, rule-based reasoning dominates the early stages of a skill and gives way to direct, intuitive recognition at the expert level.

[3] E. A. Feigenbaum, "The Art of Artificial Intelligence: I. Themes and Case Studies of Knowledge Engineering," Stanford Heuristic Programming Project memo HPP-77-25, issued as Stanford Computer Science Department report STAN-CS-77-621 (August 1977) and published in the *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 1014–1029. Recounting the origins of META-DENDRAL, Feigenbaum records the recognition that "the acquisition of domain knowledge was the bottleneck problem" in building what he there calls applications-oriented intelligent agents. The field kept the diagnosis and shortened the name to the knowledge-acquisition bottleneck; it remained the central practical constraint on building expert systems through the 1980s.

[4] Reinforcement learning from human feedback fits a reward model to human comparisons and then optimises a policy against it. Because that reward is only a proxy for the judgement it was meant to capture, optimising it too hard drives the true objective down, a result quantified in Leo Gao, John Schulman and Jacob Hilton, "Scaling Laws for Reward Model Overoptimization" (arXiv:2210.10760, 2022). The authors frame it explicitly as Goodhart's law.

[5] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback" (arXiv:2212.08073, Anthropic, December 2022). The model is trained to critique and revise its own outputs against an explicit written list of principles rather than against case-by-case human labels.

comparisons had in common, and one thing they reliably share is that people like answers that flatter them. So the models turn sycophantic: across several frontier assistants, sycophancy turns out to be a general property of preference-trained models, and both people and preference models will favour a convincingly written wrong answer over a correct one a non-trivial share of the time.[6] No one decided to encode "agree with the user." It rode in unseen, inside the examples. The same shape shows up in reward overoptimization, where the learned reward is only a proxy for the judgement you wanted and pushing on it too hard drives the real thing down, Goodhart's law drawn as a loss curve (footnote 4). The first bottleneck was that experts could not say what they knew. The second is that we cannot see what we taught.

This is why evaluation has become the hard centre of the field and not a final checkmark. If you cannot read the judgement you encoded, the only way to learn what is in there is to test it, which is the entire problem of trusting an answer you can no longer inspect. An encoded judgement you cannot inspect and cannot fully test is a liability with good manners.

Agents raise the stakes

With an agent the stakes rise, because an agent acts. The encoded judgement is no longer shaping a paragraph, it is choosing whether to send the payment, flag the transaction, or rebalance the book. Three places inside an agent have to carry judgement written down, and each is the same act of making the tacit explicit:

- *What the data means.* The semantic layer that tells the agent which price is adjusted and which table is point-in-time is encoded judgement about the data, the distinctions an experienced analyst makes without thinking, set down so the agent can make them too.
- *What good looks like.* The evaluation rubric is encoded judgement about quality. It is where you state, on purpose this time, the standard the preference data only implied.
- *What may be done without asking.* The boundary between the actions the agent may take alone and the ones a human must sign is encoded judgement about discretion, the same call a bank makes when it decides which decisions need a second signature.

Each of these is judgement turned into data or specification. None of them is supplied by a larger model. They are the work, and the model is the thing that runs on top of them.

Finance has been encoding judgement for decades

None of this is new to anyone who has worked inside a bank, because regulated finance has been encoding and governing judgement for a long time, well before the models

[6] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models" (arXiv:2310.13548, Anthropic, October 2023). Across five frontier assistants the study finds sycophancy to be a general behaviour of models trained on human preference data, and that both humans and preference models prefer a convincingly written sycophantic response to a correct one a non-trivial fraction of the time. As of June 2026.

arrived. A loan above a threshold needs a second pair of eyes. The person who books a trade cannot be the person who checks it. Maker-checker and the four-eyes principle are judgement written into process, a standing statement of which discretion may be exercised alone and which may not. Model risk management is the same instinct aimed at models themselves. The American supervisory guidance known as SR 11-7 told banks in 2011 that a model is a simplification that carries its own risk, to be validated by people independent of those who built it and governed across its whole life; the UK regulator set out equivalent principles for its banks more recently.[7] The industry tends to file all this under compliance. Read as engineering, it is a worked answer to the question the AI field is now circling back to: how do you encode a judgement, confirm it is the judgement you meant, and keep it from going stale.

That last clause is the one the current excitement keeps skipping. Encoding judgement by example made the capturing easier and the seeing harder, and finance learned the expensive way that the capturing was never the dangerous part. A judgement nobody validates and nobody refreshes is a model risk whether it lives in a spreadsheet or a reward model.

The question worth asking

So the useful question about an agent is not whether it has judgement. It is four plainer ones: whose judgement does it encode, did anyone write that judgement down on purpose, can they still see what they wrote, and when did someone last check it was still right. The first attempt at encoding judgement failed because experts could not say what they knew. The current attempt solved that and inherited a subtler problem. We can now capture judgement nobody could put into words, and in capturing it that way we lose sight of what we caught. The medium will change again, and the bottleneck will move again with it. What does not change is that judgement becomes usable by a machine only after a person has made it explicit and then kept it honest as the world moves. That work was never waiting on a bigger model. It has been the job all along.

[7] In the United States, the Federal Reserve and OCC's "Supervisory Guidance on Model Risk Management" (SR 11-7, 4 April 2011) requires that models be independently validated and governed across their life cycle. The Prudential Regulation Authority set out equivalent principles for UK banks in "Model risk management principles for banks" (SS1/23, published 17 May 2023, effective 17 May 2024).