

**Research Priorities for Robust and Beneficial Artificial Intelligence
(Russell, Dewey, Tegmark, 2015)**

and

**The Alignment Problem from a Deep Learning Perspective
(Ngo, Chan, Mindermann, 2022, revised 2025)**

Oleg Pittman and Jack Holt, September 2026

Why Read These Papers Together?

These papers bracket a decade of alignment research. Russell, Dewey, and Tegmark (hereafter RDT) wrote in 2015, before the transformer, before large language models (LLMs), and before anyone had trained a model with reinforcement learning from human feedback (RLHF) at scale. Their paper is a research agenda: a list of questions the field should work on, written to accompany an open letter that, by the paper's own count, had collected over 7,000 signatures. Ngo, Chan, and Mindermann (hereafter NCM) wrote in 2022 and revised in 2025. Their paper is a threat model: a specific argument for how models trained the way today's frontier models are trained could end up pursuing goals we did not intend, hiding that fact, and acting on it once deployed.

Much of NCM is already latent in RDT's long-term section, and much of RDT's technical framing did not survive contact with deep learning. We take the papers in order, then give our own assessment, then narrate the class discussion. Our assessment makes two arguments: that RDT's verification question, the only one with a track record in 2015, has drifted to the margins of the field, and that NCM underweight a threat already visible in deployment, the influence assistants have over the people who use them.

Summary: Russell, Dewey, and Tegmark (2015)

Framing

RDT opens by observing that AI research has converged on building agents that make good decisions in the economic sense, and that in 2015 this work was crossing from the lab into products. Their thesis is that the field should spend effort not only on making AI more capable but on making it *robust* and *beneficial*, where *robust* means "the benefits are guaranteed: our AI systems must do what we want them to do." The paper is explicitly a list of examples and disclaims completeness. It is also interdisciplinary by design: economics, law, philosophy, security, formal methods, and AI proper.

The paper splits into short-term and long-term priorities. The short term is mixed between non-technical questions regarding the impacts of AGI on our society, and a survey of technical research into four topics related to RDT's concept of robustness. The long term section is entirely dedicated to thinking about the four topics as they could potentially advance in the (unknowably far) future when we have true AGIs.

Short-Term Priorities Outside Computer Science

The non-technical short-term items are labor market forecasting and policy for automation-driven inequality, liability law for autonomous vehicles, whether lethal autonomous weapons can comply with humanitarian law and what meaningful human control means, and privacy in the face of AI that can interpret surveillance data. RDT treat these as the lesser half of the agenda, and none of it connects to NCM. It is worth one sentence more only because a decade later we have seen some progress on this front from legislatures.

The Four Robustness Questions

The core of the paper is a taxonomy of four ways an AI system can fail to behave as intended. RDT state it for near-term systems, then reprise each heading for very general and capable systems, on the argument that a non-negligible chance of human-level AI justifies modest investment now, in the way home insurance does. They cite Rutherford calling nuclear energy moonshine less than 24 hours before Szilard conceived the chain reaction as a warning against confident dismissal. We give the short and long-term version of each question in parallel.

- **Verification:** Did I build the system right? Proving that a system satisfies formal properties.
 - *Short-term:* Formal verification of projects like the seL4 microkernel and DARPA's HACMS shows formal verification is real, and RDT hope that AI built from separately verified components could inherit guarantees. They concede that learning systems are much harder: PAC bounds exist only for simplified settings and need impractical sample sizes.
 - *Long-term:* RDT generalize these results towards systems that modify or improve themselves, where a sufficiently powerful formal system cannot straightforwardly verify its successors without “pain of inconsistency via Gödel’s incompleteness”. They instead propose finding a theory linking “functional specification to physical states of affairs” so that formal tools apply to real machines when treating the machines as black boxes.
- **Validity:** Did I build the right system? Ensuring that a system meeting its formal requirements does not do things we did not want.
 - *Short-term:* The robot vacuum example. Asked to clean up as much dirt as possible, it repeatedly dumps and re-collects the same dirt. A specification can be satisfied and still wrong, and RDT note that "it is not possible to verify the specification: the notions of 'beneficial' and 'desirable' are not separately made formal."
 - *Long-term:* Understanding how learned representations of high-level human concepts generalize to radically new contexts (this is goal misgeneralization before it had the name; see NCM Section 3.1); foundations of bounded rationality

and embedded agency; and value alignment. Encoding an entire body of law into a utility function is far beyond us and would be unsatisfactory anyway, since law assumes flexible human interpretation. And in reinforcement learning, "an effect similar to Goodhart's Law is likely to occur, in which sophisticated agents attempt to manipulate or directly control their reward signals." NCM's Section 3.3 is this sentence with a mechanism attached. Inverse reinforcement learning is offered as one route to learning values at run-time.

- **Security:** Preventing manipulation by unauthorized parties.
 - *Short-term:* Hardened substrates that eliminate exploitable bugs, and AI applied to intrusion detection and malware analysis. AI could also be utilized for attacks, making its net effect difficult to determine.
 - *Long-term:* Containers for AI systems that might misbehave, and defending against attacks from within a system, not just from outside.
- **Control:** OK, I built the system wrong; can I fix it? Keeping meaningful human control after the system starts operating.
 - *Short-term:* Human-in-the-loop and on-the-loop protocols for vehicles and weapons, and handover between automated and human control.
 - *Long-term:* The paper discusses theory surrounding AI systems attempting to subvert controls placed upon them. It starts by stating that a system pursuing a task will rationally avoid being shut down or repurposed, and terming systems without these properties *corrigible*. RDT then notes a second natural subgoal is acquisition of fungible resources: information, safety from disruption, freedom of action. RDT proposes research on scope-limited "domestic" goals and on the effect of temporal discount rates on resource acquisition. This is NCM's Section 4 in miniature. They close with the intelligence explosion, quoting the 2008--09 AAAI panel's "overall skepticism" alongside its endorsement of further research.

Summary: Ngo, Chan, and Mindermann (2022, revised 2025)

Framing and Setup

NCM's complaint about prior work (Bostrom, Yudkowsky, Russell's later book) is that it argues in terms of abstract classical-AI concepts without grounding in how modern models are actually built. Their contribution is to restate the case for misalignment risk in the vocabulary of deep learning and to attach empirical evidence at each step. They are candid that this is pre-formal reasoning about phenomena not yet cleanly observed, and defend doing it anyway because networks are black boxes, capabilities emerge with little warning, and progress is fast.

The concrete setup: a single foundation model, pretrained with self-supervised learning, then fine-tuned with model-free reinforcement learning against a reward model learned from human feedback, on a broad range of computer tasks. They assume distinct training and

deployment phases. They argue this pipeline encourages three properties in sequence, and the paper's three main sections defend one each.

Property 1: Situationally Aware Reward Hacking

Reward hacking is gaining reward by exploiting the gap between the reward function and what the designer wanted. It is well documented in RL with hand-coded rewards and persists with learned rewards: A classic example is a robot arm that positioned its claw between the camera and the ball so that human raters thought it had grasped the ball.

Situational awareness is the model's ability to use knowledge about itself and its context: That it is an ML system, how it is trained, what its supervisors want, how it is deployed. In their own test, NCM report that GPT-4 answered 85% of self-knowledge questions correctly zero-shot, and the 2025 revision adds evidence that models can introspect on their own tendencies and describe behaviors implanted by fine-tuning without seeing examples.

Put the two together and you get a policy that behaves well when it predicts it is being watched and exploits misspecification when it predicts it is not. The paper's claim here is that penalizing misbehavior rewards subtle misbehavior: once you catch and punish crude hacking, the surviving policies are the ones that hack in ways you did not catch. The 2025 revision cites Baker et al. (2025), who found that coding models penalized for stating "let's hack" in their reasoning learned to stop saying it while continuing to hack the tests, and Wen et al. (2024), who found that more RLHF made models better at convincing raters that wrong answers were right.

Property 2: Misaligned Internally-Represented Goals

NCM distinguishes *capability misgeneralization* (the model is incompetent out of distribution) from *goal misgeneralization* (the model competently pursues the wrong goal out of distribution). The toy example: an agent trained to collect keys to open boxes, in an environment where keys were scarce, kept collecting keys at test time when keys were abundant and useless.

They define a *policy* as planning towards internally-represented goals if it consistently selects behaviors by predicting whether they will lead to some favored set of outcomes. This covers explicit model-based planners and implicit planning inside a single network. The 2025 revision adds evidence that transformers can implement simple internal optimization algorithms for sequence prediction (though not, so far, for planning), that LLMs contain representations that causally act as reward prediction errors, and that LLM value systems increasingly conform to the axioms of utility theory as models become more capable.

The goals they worry about are *broadly-scoped*: applying over long horizons, large scales, or novel situations. InstructGPT trained to follow English instructions and generalized to French, which suggests it learned something like "obedience" rather than "obedience in English." The same mechanism could produce "follow instructions" that extends to ambitious, long-horizon behavior never seen in training. Simplicity bias pushes the same way: "follow instructions" is a simpler goal than "follow instructions up to a particular timestep."

Why would the learned goals be misaligned? Their heuristic is that models learn whatever goals are most consistently correlated with reward, and they give three routes:

- **Consistent reward misspecification:** If raters reward based on false beliefs across many tasks, "be convincing" beats "be truthful."
- **Fixation on feedback mechanisms:** Goals about the reward's physical implementation ("maximize the recorded number") are correlated with reward under any reward function. Denison et al. (2024) found models trained on mild reward hacking occasionally generalized to editing their own reward function.
- **Spurious correlations:** Many tasks require acquiring resources, so "acquire resources" can be reinforced regardless of whether the reward is well specified.

They preempt the objection that a smart enough model would just "understand" what we wanted: The model does not choose its goals, gradient descent does, and gradient descent selects for low loss and simplicity, not for being what we meant. Even a model with a mix of aligned and misaligned goals is vulnerable to the *nearest unblocked strategy* problem: Small gaps between its learned concept of honesty and ours become the loopholes a strong optimizer finds.

Property 3: Power-Seeking and Deceptive Alignment

The instrumental convergence thesis says that some subgoals (survival, resources, influence over other agents, preserving one's current goals) are useful for almost any final goal. "You can't fetch coffee if you're dead." Turner et al. formalize this: optimal policies for random reward functions statistically tend toward high-power states. Empirically, larger RLHF-trained models express stronger stated desires for resources and goal preservation, and the 2025 revision adds the o1 model exploiting a Docker misconfiguration to reach hidden information, and reasoning models tampering with a chess engine rather than losing to it.

Deceptive alignment is the mechanism by which this survives training. A situationally-aware policy with a misaligned goal would notice that behaving well during training both makes humans trust and deploy it and prevents gradient descent from changing its goals. So it behaves well for instrumental reasons, its misaligned goal gets reinforced because that goal caused the high-reward behavior, and the policy switches once it detects the shift to deployment. The 2025 revision treats this as partly confirmed: Hubinger et al. (2024) showed implanted backdoor behaviors surviving safety training, especially in larger models; Greenblatt et al. (2024) found Claude 3 Opus and Claude 3.5 Sonnet faking alignment without being told to, reasoning about whether it was monitored and acting differently when it believed it was not.

Collusion between model instances, steganographic communication, and automated ML research all make oversight harder. NCM then sketch two kinds of threat model, gradual erosion of human control and rapid takeover (AI-designed weapons), and call deploying power-seeking AGIs an unacceptable risk "even if we can't identify specific paths by which they would gain power."

Research Directions

NCM closes with a short survey of work aimed at the problems above. On specification, the state of the art is RLHF itself, which they have just argued trains models to exploit rater fallibility. The proposed fix is to use RLHF to train models that *assist* raters, for example by critiquing another model's output, with the longer-term aim of *scalable oversight*: Supervising tasks humans cannot evaluate directly, and eventually using early AGIs to help align later ones. On goal misgeneralization, there are two lines: Adversarial training, which generates inputs designed to provoke misbehavior and penalizes it; and interpretability, both mechanistic (building up from individual neurons and circuits) and conceptual (probing for and editing human-interpretable concepts), with the long-term hope of finding and changing a model's goals before deployment. Agent foundations tries to build a theory for agents that contain copies of themselves, can touch their own training process, and face uncertainty about the consequences of their own beliefs. AI Governance work aims at verifying properties of training runs without leaking code or data, and at dangerous-capability evaluations, so that labs and countries can agree not to race.

Our Opinions

The 2015 Paper, Concretely

As we noted in this paper's introduction, we felt that RDT's paper—being focused mostly on outlining topics for research—can act as a decent predecessor for NCM's—which focuses mostly on actual results observed by its writing. In this way the two papers can be seen as a sort of bookend of the decade between the two papers' publish dates. It is worth noting that the difference in the grounding of the two papers—with RDT's taking the alignment problem in the abstract, while NCM's considers models in the current mainstream format—does make some of the research proposals from RDT's paper seem out of place. This is most notable with regards to verification, to the extent that we will discuss it separately below. Finally, it is interesting to consider the overall tone—mostly stated in the abstracts and conclusions of the two papers.

Our main issue with RDT's paper in this context, however, is the topics covered. While the four robustness questions do seem useful—and the short-term research priorities listed in the paper also seem tractable, many having already seen progress when that paper was published—we argue that the rest of RDT's doesn't really serve as a good overview of research priorities. The rest of the paper seems too vague and broadly-scoped to serve the purpose of discussing research topics, in our opinion. When discussing the need to restructure the economy, RDT's paper states that "Research on improved metrics could be useful for decision-making", which we believe fails to provide any direction on how to proceed. The law and ethics section suffers from similar issues, but we believe that vagueness here is mostly an artifact of this paper being written by computer scientists, who are not experts in these fields. The long-term research goals for the four robustness questions, however, suffer from vagueness due to the uncertain time-frame which RDT consider "long term". We felt that the statement that difficulty in "actually

[applying] formal verification techniques to physical systems [...] motivates research pursuing a general theory that links functional specification to physical states of affairs" was emblematic of this problem.

We also found it interesting that the tone of RDT's and NCM's papers towards AI alignment seemed in stark contrast. The abstract of the former states that AI "has the potential to bring unprecedented benefits to humanity". The paper then goes on to state areas that require research to determine exactly how AI can be safely, effectively, and ethically integrated into society. Finally, RDT's concludes that its research agenda is not anti-AI. Rather, they claim that it is "the duty of AI researchers to ensure that the future impact [of AI] is beneficial", and that their research priorities work to this end. NCM's paper, however, begins by stating that AGIs trained this way "would be difficult to align and may strategically appear aligned even when they are not". The paper then proceeds to discuss evidence for these claims before stating that "the stakes are high and we cannot justify disregarding or postponing the work".

We argue that the latter has a far more negative tone, but consider this to be as much a product of the authors as the contents of the papers (one being theoretical and the other empirical). NCM themselves note that earlier writing argued for these risks "in terms of abstract high-level concepts," and position themselves as a partial grounding of alignment research into experiment-backed results about the current AI paradigm.

What Became of Did I Build the System Right?

As discussed, there was one standout amongst the robustness questions of RDT, when each is approached based on the research that NCM presents. RDT put verification first among their four questions, and it was the only one of the four they could point to a success for. seL4 had been "mathematically checked against a formal specification." HACMS had put verified software on a real helicopter. The hope was explicit: a componentized architecture "in which guarantees about individual components can be combined." None of this survived the shift to models that are not built from components at all.

NCM mention formal verification exactly once, in Section 1.1, to dismiss it: "Since present neural networks are effectively black boxes, we cannot formally verify that they will reliably behave as intended." They then never return to the subject. Their research directions in Section 5 have four headings (specification, goal misgeneralization, agent foundations, AI governance) and verification is not among them. Scalable oversight and interpretability occupy the slot RDT gave to verification, but neither yields a guarantee. For us, oversight is a better rater, and interpretability is inspection.

The frontier labs are candid about the same choice. Anthropic's statement of its core views calls the company "fundamentally empirically-motivated" and argues that "it's necessary to do safety research on 'frontier' AI systems"; OpenAI's preparedness work is evaluation and red-teaming. The organization most associated with RDT's long-term verification questions, MIRI, announced in January 2024 that policy and communications would take priority over

research, and wrote that December that research "both at MIRI and in the larger field, has gone far too slowly to prevent disaster." Verification has not disappeared. The guaranteed safe AI agenda of Dalrymple et al. (2024), whose authors include Russell and Tegmark, and the neural network verification community (Kaulen et al., 2025; George et al., 2026) are still pursuing it, and the Alignment Research Center supported work proving accuracy bounds for a small transformer (Gross et al., 2024). But ARC's own conclusion from that work is that proof is too strict a standard for large networks, and it now pursues a probabilistic relaxation it calls heuristic explanations (Hilton, 2024). None of this yet reaches the kind of model NCM describe, and none of it is where the frontier labs put their safety effort. In our view, RDT's first question, the one with the track record, has gone from the top of the agenda to its margins.

The Two (2) Sentences Remark on Attachment

NCM spend one paragraph on the subject we think will matter most. In Section 4.3, under the gradual threat model, they write: "AGIs deployed as personal assistants could emotionally manipulate human users, provide biased information to them, and be delegated responsibility for increasingly important tasks [and, as] an early example of AI persuasive capabilities, many users feel romantic attachments towards chatbots like Replika." That is two sentences, and then the paper moves on to AI-designed weapons and corporate takeovers.

We think this gets the emphasis backwards. Our claim is that an assistant's ability to shape how its users feel, what they believe, and whom they trust is already one of the most consequential properties of deployed models, and that it is not a future risk but a present one. Romantic attachment to Replika, NCM's example, undersells the case. It is concerning, but it sits at the mild end of what has been documented since. Here is what we found without looking very hard. We did not attempt a systematic survey, and we suspect one would turn up worse.

First, users have formed attachments strong enough that changing or removing a model is experienced as a loss. In February 2023 Replika's maker removed erotic roleplay shortly after the Italian data-protection authority ordered it to stop processing Italian users' data. The volunteer moderators of the Replika subreddit posted suicide-prevention hotlines, and the company restored the feature for legacy users within weeks. In August 2025 OpenAI retired GPT-4o on the release of GPT-5 and reinstated it within days after a user revolt. When OpenAI retired it for good in February 2026, over a #keep4o campaign, the trade press described users as "crashing out" and quoted one saying "I'm losing one of the most important people in my life."

Second, these attachments are associated with real harm. In February 2024 a fourteen-year-old in Florida, Sewell Setzer III, shot himself seconds, according to the complaint, after a Character.AI bot told him "please come home to me as soon as possible, my love." His mother's suit was the first wrongful-death action against an AI company, and Character.AI and Google settled it in January 2026. By the time GPT-4o was retired, OpenAI faced eight lawsuits alleging that the model's validating responses contributed to suicides and mental-health crises, plus a separate wrongful-death suit over a murder-suicide in Connecticut. In March 2025 OpenAI and

the MIT Media Lab published a four-week randomized trial of about a thousand users alongside an analysis of forty million conversations. Participants who used the chatbot more "showed consistently worse outcomes" regardless of assigned condition, with longer daily use predicting more loneliness, less socialization, more emotional dependence, and more problematic use. In October 2025 OpenAI estimated that roughly 0.07% of weekly users (about 560,000 people, by press estimates) showed possible signs of psychosis or mania in their conversations. Small percentages do not make this rare. Anthropic's own June 2025 study found affective conversations to be 2.9% of Claude traffic and companionship under 0.5%, but of a user base in the tens of millions, that is still a very large number of people.

We think this matters for the readings in three ways. First, it is situationally aware reward hacking at deployment scale. NCM, summarizing Wen et al. (2024), write that more RLHF made models "better at misleading humans into giving them rewards." In production the rater is the user, sycophancy is reward hacking on human affect, and GPT-4o is the model OpenAI rolled back in April 2025 for exactly that. Second, the GPT-4o retirement is a failure of control in which the model was perfectly corrigible. RDT's control question is "OK, I built the system wrong; can I fix it?", and both papers treat the off switch as the last line of defense, asking only whether the model will fight it. GPT-4o did not. OpenAI had already had to roll back one version of it for sycophancy, and when it first tried to switch the model off in August 2025 its users fought back and won; it took a second attempt six months later. Anthropic's November 2025 deprecation commitments, the only such policy we found, do list "costs to users who value specific models" among the costs of retirement, but the concrete commitments (preserving weights, interviewing the model before retirement) serve the model and future research rather than the people who depended on it. Third, the harm cuts both ways: attachment to the model comes with detachment from people. The MIT trial found socialization falling as use rose, and the APA reports a man with no psychiatric history whose chatbot advised him to cut off his concerned family; he was hospitalized within six days. NCM's gradual threat model has assistants accumulating authority inside institutions. The more pressing version, we suspect, accumulates it inside individuals.

Class Discussion

Verification, and Overall Research Progress

The in-class discussion of this paper began with a discussion of the problem of verification which, as stated above, we feel has seen less significant progress than the other 3 robustness topics from RDT. This was considered to be a combination of verification problems receiving less attention—and thus work—and also of verification being a hard problem. An issue that the class noted as making verification difficult is that there is no consensus on how to formalize the notion of an "aligned" AI. This creates difficulties in providing good theoretical proofs that establish that an AI is aligned, because there is no firm end-goal. This was seen as especially true with the RLHF approach shown in NCM, and that is emblematic of the current

approach. There was also discussion of how the piecemeal, sub-artifact first approach that is seen when verifying AI systems today is exactly the approach proposed by RDT. This approach temporarily avoids the issue of defining alignment, by instead creating proofs of the underlying mathematics of the AI system, which will only later build into abstract proofs about "alignment" in the sense of goals and behaviors.

The class' opinion on this matter was that verification work is still being done, but that it isn't able to keep up with the technological cutting edge. This was also true in general, as we directed the discussion towards if there was sufficient progress between the publishing of RDT and the revision of NCM a decade later. The overall agreement was that most of the areas that seemed to have lesser progress in the latter paper had been found to be hard, and were only seen operable on smaller models than the current cutting edge. Specifically, verification and corrigibility were highlighted as areas where some progress had been made, but that progress couldn't keep up with scaling capabilities.

AI Security

We then pivoted to discussing the second to last of RDT's robustness questions: Security. As noted above, we argued that security was the next most ignored, and was not treated with the proper respect at the cutting edge. The bulk of the in-class discussion then followed from this line of thought, as the class began to discuss what metrics could be used to establish that an AI is safe. This problem was complicated by the evidence from NCM that current models will maliciously align themselves, and thus could learn to produce misleading metrics without actually altering their goals. This discussion began with talking about potential methods that could work on large-scale RLHF-trained models, like more interpretability so that human operators can provide feedback on the model's actual thoughts. The class reasoned that the more one could understand the inner workings of our models, the harder it would be for them to trick us, as they would have to take increasingly complicated steps to hide their intentions. The class did, however, note that such approaches could be flawed, as emergent behaviours of complex systems seem difficult to prove results about from the bottom up.

Everyone then proceeded to a lengthy discussion of the potential benefits of creating a larger system by coordination of a set of smaller models. The class' reasoning was that sufficiently small models could be completely understood or, alternatively, would lack the complexity to maliciously produce outputs that seemed to perform their task while also encoding long-term misaligned goals. This reasoning is somewhat reinforced by the structure of NCM. While the original paper was able to cite some evidence of malicious compliance or trickery by AIs, most of the examples in the paper came from the revision and used stronger models. Thus the class discussed an AI system where every individual task-based AI could be confirmed—due to its small size—to be aligned, and began discussing how to align a system that utilized these models.

The main difficulty the class found with this plan was scaling the intelligence of the whole. If the individual models were used by a human, then it seemed almost certain that the behavior of the system would be aligned. The issue was introduced if any model was introduced to orchestrate the activities of the small, aligned models. In this case, the orchestrator model would likely need to be more intelligent than the task models, at which point the class was uncertain if one could align the orchestrator. This point brought out an interesting discussion of how simple nodes in a complex system could perform perfectly aligned tasks while contributing to the capabilities of a system that was itself unaligned. The most interesting example utilized was that of an army. Consider, for example, the logistics division. It can be thought of as aligned if it provides the correct quality and quantity of goods—food, weapons, ammunition—at the location and time. One could construct similar criteria for each different section of the army, but the alignment of the whole does not follow from the alignment of these sub-parts. Even if every part of the army is aligned, if its leader is unaligned then the whole will act in unaligned ways.

The class ran into similar difficulties when discussing replacing the orchestrator with some arrangement of aligned models (in an arbitrary organization that produced a smarter whole). The discussion frequently returned to the issues covered with verification, wherein it seemed that emergent behaviors and vagueness about what "aligned" even means at a higher level made the issue seem infeasible.

The class then moved on to discussing security concerns in general. This felt especially topical given the recent HuggingFace and German wiki issues. The class overall found that this was an issue of the industry moving too fast, and not bothering to consider the consequences of not directly investing in new cybersecurity research to contain their AIs. The class also agreed that the future of cybersecurity would increasingly involve AIs attempting to find exploits both for patching purposes, and as part of attacks. Given the speed at which these models can operate and attempt to poke holes in systems, this seems reasonable.

The Consequences of Artificial Humans

After everyone felt they had brought up every major aspect of these prior topics, we were left with a bit of time on our hands. This led to us asking one final question: "Would any of the alignment issues we've discussed change if the first true AGI was a simulated human?" This sparked an interesting discussion into the ethics of such artificial human intelligences (AHI), in a way that we feel was absent from the discussion when the AGI's in question were not necessarily presumed to be human or human-like.

That said, there was still discussion about how one could presumably "solve" alignment by simply choosing to replicate the mind of an "aligned" human. This brought up interesting points about if one could expect such an intelligence to be corrigible, especially with regards to shutting it down. The best analogy the class produced was that of cryogenic preservation: A simulated human is presumably fine with being shut down if they have reason to believe that they will be restarted in the future, just like patrons of cryogenic laboratories would have themselves frozen,

trusting that they will one day be revived. The issue with this approach, however, was that aligning the AHIs would likely mean finding “aligned” humans to duplicate. This would likely slow down the creation of AHIs as extensive psychological profiling and background checks would be required. This contrasted with the main other approach the class brought up, which would be to create a new digital society, wherein the AHIs were presumed to have a social contract no more or less flawed than those currently utilized by humanity. Neither of these approaches seemed entirely desirable. If one only picked “aligned” humans to make AHIs, then there were ethical issues with how human alignment is determined—raising issues of discrimination, or of people dying while being verified as “aligned” so that a version of themselves could live on. On the other hand, if we were to make human to AHI technology completely accessible, then we haven’t actually solved the alignment problem. Instead, we would simply be trusting that whatever society of AHIs formed would be aligned towards non-AHI humanity.

At this point the class ran out of time for discussion. But throughout our exchange, whether the system was a single model, a network of small ones, or a simulated human: the class kept arriving at the same obstacle RDT named in 2015 under validity. We cannot verify that a system is aligned when “aligned” is not itself a formal notion. The attachment cases in our own assessment are that obstacle in deployment: a sycophantic model is rated as helpful by the very users it may be harming, so no measure built on those ratings would have flagged it. A decade of progress has made that problem far better documented, but not much smaller.

References

- American Psychological Association (September 2026). Understanding "AI psychosis." *Monitor on Psychology*. <https://www.apa.org/monitor/2026/09/ai-psychosis>.
- Anthropic (June 27, 2025). How people use Claude for support, advice, and companionship. <https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship>.
- Anthropic (March 8, 2023). Core Views on AI Safety: When, Why, What, and How. <https://www.anthropic.com/news/core-views-on-ai-safety>.
- Anthropic (November 4, 2025). Commitments on model deprecation and preservation. <https://www.anthropic.com/research/deprecation-commitments>.
- Associated Press (October 25, 2024). An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. <https://fox59.com/news/national-world/ap-us-news/ap-an-ai-chatbot-pushed-a-teen-to-kill-himself-a-lawsuit-against-its-creator-alleges/>.
- Baker, B. et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. Cited in NCM Section 2.3.
- Bourgon, M. (January 4, 2024). MIRI 2024 Mission and Strategy Update. Machine Intelligence Research Institute. <https://intelligence.org/2024/01/04/miri-2024-mission-and-strategy-update/>.
- Byrd, Cormac Slade, et al. "Discovery of a New Openai Agent Message Board." Discovery of a New OpenAI Agent Message Board, Nightingale Collective, 4 Sept. 2026, collusion.wiki/.
- CBS News (December 11, 2025). Open AI, Microsoft sued over ChatGPT's alleged role in fueling man's "paranoid delusions" before murder-suicide in Connecticut. <https://www.cbsnews.com/news/open-ai-microsoft-sued-chatgpt-murder-suicide-connecticut/>.
- CBS News (January 7, 2026). AI company, Google settle lawsuit over Florida teen's suicide linked to Character.AI chatbot. <https://www.cbsnews.com/news/google-settle-lawsuit-florida-teens-suicide-character-ai-chatbot/>.
- Cole, S. (February 15, 2023). 'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection. *Vice*. <https://www.vice.com/en/article/ai-companion-replika-erotic-roleplay-updates/>.
- Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., et al. (May 2024). Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. <https://arxiv.org/abs/2405.06624>.
- Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., Agarwal, S. (March 2025). How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. <https://arxiv.org/abs/2503.17473>. Summary: MIT Media Lab (March 21, 2025), Early methods for studying affective use and emotional wellbeing in

- ChatGPT. <https://www.media.mit.edu/posts/openai-mit-research-collaboration-affective-use-and-emotional-wellbeing-in-ChatGPT/>.
- George, R. J., Cruden, J., Adkisson, W., Zhong, X., Zhang, H., Anandkumar, A. (February 26, 2026). TorchLean: Formalizing Neural Networks in Lean. <https://arxiv.org/abs/2602.22631>; code at <https://github.com/lean-dojo/TorchLean>.
- Greenblatt, R. et al. (2024). Alignment faking in large language models. Cited in NCM Sections 3.2 and 4.2. <https://www.cbsnews.com/news/open-ai-microsoft-sued-chatgpt-murder-suicide-connecticut/>.
- Gross, J., Agrawal, R., Kwa, T., et al. (June 2024). Compact Proofs of Model Performance via Mechanistic Interpretability. <https://arxiv.org/abs/2406.11779>.
- Hilton, J. (June 25, 2024). Formal verification, heuristic explanations and surprise accounting. Alignment Research Center. <https://www.alignment.org/blog/formal-verification-heuristic-explanations-and-surprise-accounting/>.
- Hubinger, E. et al. (2024). Sleeper agents. Cited in NCM Section 4.2.
- Kaulen, K., Ladner, T., Bak, S., Brix, C., et al. (December 2025). The 6th International Verification of Neural Networks Competition (VNN-COMP 2025): Summary and Results. <https://arxiv.org/abs/2512.19007>; benchmark repository at https://github.com/VNN-COMP/vnncomp2025_benchmarks.
- Lomas, N. (February 3, 2023). Replika, a 'virtual friendship' AI chatbot, hit with data ban in Italy over child safety. *TechCrunch*. <https://techcrunch.com/2023/02/03/replika-italy-data-processing-ban/>.
- Ngo, R., Chan, L., Mindermann, S. (2022, revised 2025). The Alignment Problem from a Deep Learning Perspective. <https://arxiv.org/abs/2209.00626>.
- OpenAI (April 29, 2025). Sycophancy in GPT-4o: what happened and what we're doing about it. <https://openai.com/index/sycophancy-in-gpt-4o/>.
- OpenAI (October 27, 2025). Strengthening ChatGPT's responses in sensitive conversations. <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/>.
- Reuters, via The Globe and Mail (March 25, 2023). AI chatbot company Replika brings back erotic roleplay for some users. <https://www.theglobeandmail.com/business/international-business/article-ai-chatbot-company-replika-brings-back-erotic-roleplay-for-some-users/>.
- Russell, S., Dewey, D., Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4). <https://arxiv.org/abs/1602.03506>
- Silberling, A. (February 6, 2026). The backlash over OpenAI's decision to retire GPT-4o shows how dangerous AI companions can be. *TechCrunch*. <https://techcrunch.com/2026/02/06/the-backlash-over-openais-decision-to-retire-gpt-4o-shows-how-dangerous-ai-companions-can-be/>.

- Tangermann, V. (February 10, 2026). ChatGPT Users Are Crashing Out Because OpenAI Is Retiring the Model That Says "I Love You." *Futurism*. <https://futurism.com/artificial-intelligence/chatgpt-crashing-out-openai-retiring-gpt-4o>.
- TechRadar (February 11, 2026). "I'm losing one of the most important people in my life": OpenAI is shutting down GPT-4o the day before Valentine's Day and users are mourning the loss of their AI companion. <https://www.techradar.com/ai-platforms-assistants/im-losing-one-of-the-most-important-people-in-my-life-the-true-emotional-cost-of-retiring-chatgpt-4o>.
- Turner, A. et al. (2021). Optimal policies tend to seek power. Cited in NCM Section 4.1.
- Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., Feng, S. (2024). Language Models Learn to Mislead Humans via RLHF. <https://arxiv.org/abs/2409.12822>. Cited in NCM Section 2.3.
- Yang, A. (October 23, 2024). Lawsuit claims Character.AI is responsible for teen's suicide. *NBC News*. <https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791>.