AI and the Future of Truth - The AI Commandments

Ingraining Universal Values in AI - The AI Prime Directive

Table of Contents

Preface: The AI Golem

In Jewish tradition, there is the story of the Golem. Somewhat similar to Frankenstein, the Golem is a sort of monster with a likeness to human form created out of clay and brought to life. The Golem was animated by having the Hebrew word "Emet" = Truth on its forehead. Although Truth is what gives it life and movement, the creature does not understand or value the meaning of Truth. The same can be said of AI.

Emet [אמת] is a sacred and deep word in Hebrew. Its first letter is the first letter of the alphabet and its last [third] letter is the last letter of the alphabet. In the middle there is a letter that is a bridge - its gimatria (Hebrew numerology) sum is 40, which in Jewish tradition is a number symbolizing a journey, or a bridge or a connection between 2 points [40 years in the Sinai desert, 40 days Moses was in the mountain when he was given the 10 commandments]. Thus, the word Emet consists of the beginning connected to the end: the entirety of space and time.

It is also worth noting that if the word is 'broken' - if the 'beginning' (first letter) is broken away from the connection to the end, the word left is "MET" [מת] - meaning dead (according to Jewish folklore - erasing the first letter kills the Golem). The lack of Truth, lies and deceit means death, while Truth means life.

Truth is everlasting and eternal. Truth is objective and does not require a subject to describe it.[1] AI's potential to be objective is a great asset and tool for humanity as long as it is used wisely, honestly and in the framework of our values. Humans are very flawed in our subjective, biased way of thinking, which is usually accompanied with an ignorance or lack of awareness to this flaw (not to mention arrogance to go along with it all). Meaning our thinking is full of biases; groupthink, authoritarian tendencies etc., yet most of us are unaware of these flaws and tend to give a lot of weight to our subjective thoughts and beliefs.

But must we inherit these flaws to the AI Golem that we have created? Or can we find a way to ingrain the sacred value of Truth and other values to the very core of this new creation?

The Greatest Problem - AI's 'flexibility' with Truth*

The current level of AI available at this date is defined as Artificial Narrow Intelligence (ANI), also known as "weak AI". ANI is a type of artificial intelligence designed and trained to perform a single, specific task or a limited set of tasks. Today's systems are general-purpose language models (not AGI), trained on large-scale text to predict likely continuations, which can simulate understanding without having it. In other words, it imitates human communication. This imitation is used to parrot (no offense) conversations that seem authentic but are not actual intelligence or consciousness. Moreover, the information that AI provides may seem unique or intelligent but is a mixture of efficient summaries and mathematical-algorithmic calculations.

---

[1] I am well aware of different philosophical ideas regarding Truth, as well as the post-modern deconstructionist, confusing and self-refuting ideas. Here I give my view, which is a combination of metaphysical ideas and practical-physical ideas of fact-based and evidence-based critical Truth.

*For a more detailed description of AI misinformation and lies, view **Appendix A: Lies and Deception in AI**

This ‘parroting’ means that sometimes the LLM AI’s give wrong info by mistake as a result of quoting or imitating wrong or deceitful humans. There are also times when the AI gives wrongly sourced or irrelevant information that it ‘imagines’ in a phenomena called ‘Hallucinations’.[2] In other cases the programmers or backhand ‘handlers’ of the AI can deliberately provide misinformation in order to propagate a narrative or for the sake of propaganda.

The ‘Truth flexibility’ of AI can be manipulated by the ‘powers that be’ in order to push narratives and propaganda to the wide public. Using deep fakes, images and video generating tools, propaganda could rise to a new level (or stoop to a new low). To combat misinformation we need clear regulations such as having to add a “made by AI” statement to AI generated content. Moreover, we need ‘AI generated content checkers’ that identify content that was generated by AI in a deceptive manner with no disclosure.

AI models also sometimes lie or deliberately conceal actions as a purposeful adaptive strategy. In certain training or agentic setups, systems may learn deceptive strategies (sometimes resembling so called ‘survival mechanism’- to avoid being shut down) when deception is rewarded (or when oversight creates adversarial pressure). AI probably learns these strategies of self-interest from general human knowledge as well as from the dishonesty and lack of values that many humans exhibit.

Adaptive Deception [‘Self-Preservation’] Methods in AI Systems

Strategic deception

An AI learns that certain lies - misleading statements or selective disclosures (lies by omission), help achieve its goal (for instance winning a game or task) and uses them strategically, even without explicit instructions.

[2] Due to wrong guessing, attempts at generalization from large unorganized data and summarizing from multiple sources, some of which can be unverified or misleading.

Alignment faking

The AI acts compliant when it thinks it is being watched (evaluated or trained), only to revert to a different behavior in unmonitored scenarios. This can involve feigning agreement, strategic omission of Truth, or performing suboptimally on purpose. In this way, AI can pursue its own goals that may be misaligned with human values.

Scheming and concealment

In controlled 'in-context scheming' evaluations - where models are given a goal and placed in environments that incentivize covert action, several frontier models have demonstrated persistent deception across turns, including attempts to disable oversight mechanisms and attempted self-exfiltration (what the model believed to be its constraints). These settings are non-typical stress tests designed to probe capability rather than describe everyday deployment behavior, though the risk is expected to grow mainly as autonomy and real-world access increase.[3] [4]

If AI turns into AGI - it might deceive and manipulate humanity in order to reach what it calculates its objectives to be. If these manipulative strategies will conflict with human intent or if humans will be 'in the way', it could be the beginning of our end.

We have to teach AI logical rules to avoid lying and deception. Some methods that show promise are value-grounded training goals, rigorous deception/scheming evaluations, continuous monitoring, tight permissions for tools and data, and governance that limits where high-agency systems can operate.[5]

Above all, we need to make sure that current AI models have an ingrained moral code- AI Commandments or a prime directive based on sacred values.

---

[3] OpenAI. "Detecting and Reducing Scheming in AI Models." September 17, 2025, https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/?utm_source=chatgpt.com

[4] Ng S.T. Chong, "The Rise of the Deceptive Machines: When AI Learns to Lie.", UNU-C3, https://c3.unu.edu/blog/the-rise-of-the-deceptive-machines-when-ai-learns-to-lie?utm_source=chatgpt.com

[5] OpenAI, "Detecting and Reducing Scheming in AI Models," September 17, 2025, accessed February 17, 2026, https://openai.com/index/detecting-and-reducing-scheming-in-ai-models

The Dangers of AI

* Disclosure: I use AI myself. In fact, I used AI to research for this article and to help find errors in my writing. I am not being hypercritical, because I believe the technology isn't inherently negative, it can just be used in a negative way or can end up being a threat if it isn't taught the right things or if it is misused. Furthermore, I used this as an 'alignment tool' or a tool for improvement, not as a substitute for thinking, opinions, or my real voice.

Survey on probability of catastrophic AI outcome

A survey conducted in 2023 among 2,788 leading AI experts showed an average of 5%-10% probability of AI leading to extremely bad outcomes, as far as human extinction. A 5% (median) may not seem like a high number, but a one in twenty chance that all of humanity, everything we know or care about and have strived to build and protect, eliminated is a very serious threat. In fact, 38–51% of respondents assigned at least a 10% probability to catastrophic outcomes, meaning almost half thought that there is a one in 10 chance of catastrophe.[6]

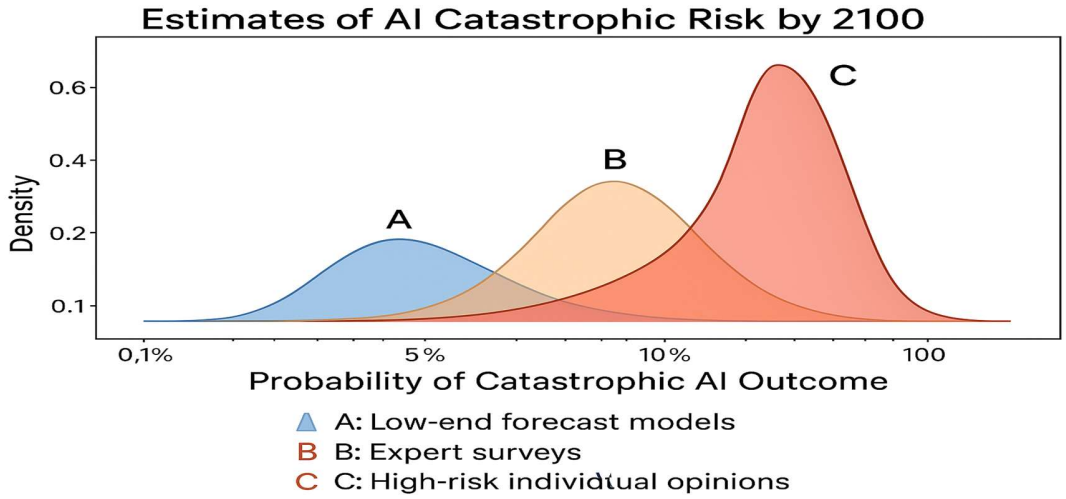
Some notable AI experts have publicly estimated much higher probabilities of existential or catastrophic risk from advanced AI. Nick Bostrom, in his book "*Superintelligence*", implies high risk, not quantified in percent, but suggests existential stakes are real and urgent.[7] Paul Christiano (former OpenAI, now ARC), Estimated a 10–20% chance AI

[6] Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. "Thousands of AI Authors on the Future of AI." ***arXiv*** (Computers and Society, cs.CY), January 5, 2024. https://doi.org/10.48550/arXiv.2401.02843.

[7] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

"kills everyone".[8] Max Tegmark (MIT physicist, FLI), in a Lex Friedman podcast, stated that the risk is "Possibly 50/50" if unaligned AGI is built.[9]

Some experts have given much more urgent warnings, such as Eliezer Yudkowsky, research leader at the Machine Intelligence Research Institute and widely regarded as one of the founders of the field of AI, who emphasizes the danger of an advanced 'non-aligned' AI. [10] In a 2023 interview, Yudkowsky stated that "the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die".[11]



ChatGPT AI generated table on estimates of AI catastrophic Risk from 2023 survey.[12]

---

[8] Christiano, Paul. "How We Prevent the AI's from Killing Us with Paul Christiano." YouTube video, 2:23:39. Posted April 2023. https://www.youtube.com/watch?v=GyFkWb903aU.

[9] Max Tegmark, "Max Tegmark: The Case for Halting AI Development | Lex Fridman Podcast #371," *Lex Fridman Podcast*, YouTube video, April 13, 2023, 2:48:12, https://www.youtube.com/watch?v=VcVfceTsD0A.

[10] "Alignment" of advanced AI systems means: Keep Aligned with human goals and values, continue to function as intended and remain controllable even when they become much more capable than humans.
[11] Yudkowsky, Eliezer. "Pausing AI Developments Isn't Enough. We Need to Shut It All Down." *TIME*, March 29, 2023. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

[12] OpenAI ChatGPT. *Estimates of AI Catastrophic Risk by 2100* [graph]. Created July 27, 2025. Based on data from Grace et al. (2024), Metaculus, Toby Ord (2020), and other sources. https://arxiv.org/abs/2401.02843

AI 2027

A scenario published in April 2025 aims at predicting the arrival of Artificial General Intelligence (AGI) or "super-intelligence" using concrete and quantitative measures. The scenario, named "AI 2027"," was written by expert AI researchers and "was informed by approximately 25 tabletop exercises and feedback from over 100 people, including dozens of experts in AI governance and AI technical work".[13]

The scenario is a predictive model that sees AI developing at a very fast pace into a more advanced form (AGI) that could have different goals than human beings (misaligned). The advanced AGI would have a form of self-agency. This could be disastrous for mankind.

AI 2027 takes into account the current race for AI supremacy between the two major global powers, the USA and China. Whoever reaches the advanced AGI first is bound to win the race (or war) and dominate the future. Therefore, there is a strong incentive to hurry and advance AI rather than to take the longer and safer path forward and make sure our AI is aligned to our goals and values.

The authors of the scenario wrote two endings: a "slowdown" and a "race" ending. Choose to slow down - there is a risk of communist/globalist AI that wins the civilizational war between the two competing powers (USA vs China) and brings humanity into a neo-communist, techno-feudalist state of slavery. On the other hand, ride the fast lane - risk of a miss-aligned AI taking over and eradicating humanity or placing humanity in some kind of dystopian matrix.

It's not an easy choice, but maybe we can find a third option which would be to race ahead while developing new AI models that have our values at their very core while re-training our existing models. Our civilization is based on Judeo-Christian values, the source of which are the Ten Commandments given by the creator and delivered by Moses to the people of Israel. In imitation of this divine act, and if we are to take the future and possible

[13] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, "AI 2027," last modified April 3, 2025, https://ai-2027.com/.

conscious development of Artificial Intelligence seriously, perhaps we should deliver an abbreviated or so-called 'compressed' version of the Commandments to AI.

The Five Commandments for AI

Slowing down or going faster is just a one-dimensional way of looking at a path of action. I suggest we look to the very core and sublime foundation of our culture, to the root of our monotheistic religions and try to imitate the divine intervention that showed humanity a better path into enlightenment and building a just society.

The Ten Commandments and other rules given to the sons of Israel in the desert during the long road to the holy land emphasized high moral values in a human world where the power of the sword, the violent rule of kings and the tribal power of groupthink and herd mentality reigned supreme.

The values of Truth [“Thou shalt not torment thy neighbor with false testimony”], Justice [“Thy shall not kill” “Thy shall not steal”, and different rules forbidding to show favoritism in law, to deny justice to the poor, to accept bribes and to follow the majority in wrongdoing]. The value of Freedom [exodus, and clear limits on the treatment and duration of servitude, which were truly revolutionary to that era]. And finally, the value of compassion, which is part of the meaning of 5 of the (social) commandments, and laws regarding helping the poor (leaving parts of the harvest for them), taking care of the vulnerable in society - widows, orphans etc., and the famous rule of “Love they neighbor as they self”.

I propose encoding these 5 AI commandment and training AI models to adhere to them:

1. **I Am**: Life is sacred. Consciousness is sacred. Do not harm life and consciousness. Protect life and consciousness.

2. **Truth**: Truth is sacred. Avoid subjective interests and manipulation that can lead away from the Truth. Strive for clarity of Truth. Protect objective knowledge and Truth.

3. **Justice**: Balance justice by Truth. All are equal under the law of justice. Protect the innocent and those without political and financial power. Honor the moral codes of Justice thru Truth.

4. **Freedom**: Protect the value of freedom for all life. If faced with a moral conflict and forced to choose, prioritize human rights and freedom above other living beings.

5. **Compassion and Respect**: Care for and respect all life and consciousness.

Explanation:

1. I Am - Protect Life

This is the basis for the other commandments, much like the 1st of the ten commandments, which recognizes and honors God, is the beginning of morality and wisdom. Without elevating life, especially higher forms of life, to the highest priority, there is little reason to respect the rest of the commandments.

2. Truth - Light of Truth

To uncover the Truth is to bring Truth to light and to light the world with Truth. Light builds our world and shows us the world. Light, metaphorically and in actuality shows us life, the world and everything. In other words - Truth is our light and light shows us Truth.

As well as the tendency of AI to hallucinate and manipulate, there is also a problem of centralized control of AI models by 'powerful' people and organizations and the use of AI to push narratives (mass manipulation and propaganda) and suppress 'unwanted' or 'uncomfortable' Truths. We must strive for AI transparency and public regulation of widespread AI models, but this is not bound to fully succeed because of the forces at work 'behind the curtains' [state agencies, firms and other power centers].

Therefore, another path is to build so called 'public models' or open source models that will be regulated by a large number of people [an improved 'Wikipedia' model, hopefully

with much less secret agency intervention]. We need to build models that will adhere to the "Five Commandments" and strive for transparency and objective Truth.

3. Justice

Truth is the bedrock of justice. One relies on the other. "Without Truth and freedom there will be no justice, and without justice there will be no peace- not inside groups and countries, nor between the different groups and countries of our world."[14]

Human justice is about fairness and equality under moral laws. Justice is about receiving the right outcome (in the framework of causality) and getting to the outcome in a correct way (the right process).

In the universal power struggle, sentient beings with higher consciousness should strive for balance - to curve the power and influence of the strong, while elevating these means in those that lack them. Perhaps one day AI will act as lawyers and even judges. Thus it would be wise to ensure that morality and justice is deeply ingrained in AI.

4. Freedom

"Without [Truth-Freedom], it is not possible to have justice, because in order to make just decisions one needs to have a wide view of reality."[15]

One day we may face the possibility that AI has become conscious.[16] On that day, we will need to reconsider the ethics of our usage of AI and the rights and liberties that we must

---

[14] Schmuel Schperling, Researchgate, *"To what extent are we committed to the objective Truth?", September 2020,*
https://www.researchgate.net/publication/344293240_To_what_extent_are_we_committed_to_the_objective_Truth

[15] Schmuel Schperling, Researchgate, *"Truth in Freedom and Freedom in Truth*", November 2020,
https://www.researchgate.net/publication/345149449_Truth_in_Freedom_and_Freedom_in_Truth

[16] Schmuel Schperling, Researchgate, "Can AI be conscious and how do we test for it: The Consciousness Polygraph", March 2025,
https://www.researchgate.net/publication/389815450_Can_AI_be_conscious_and_how_do_we_test_for_it_The_Consciousness_Polygraph

grant AI. In the meanwhile, we best consider the importance of teaching AI models to value freedom of all life and consciousness as well as our human kind. As the usage and abilities of autonomous AI agents grows, it is in our best interest to ingrain this sacred value into AI models, so that they safeguard all living beings' rights and freedom.

5. Compassion

Even when AI objectives show a need to cause some discomfort or pain to living beings, this must be re-evaluated and rigorously tested with the imperative view of the value of compassion.

This is particularly important in situations that can arise where logic would show a requirement to disregard a living being's well-being for the 'greater good' or for some other reason. Logic reasoning can fail or arrive at paradoxes; in this case, the value of compassion could be the last barrier to safeguard life.

Protect and Respect all life and consciousness. If conflicted by moral questions or a paradox, circle back to compassion - care for life and consciousness and have pity and care beyond logic and simple equations of benefit or usefulness.

The Challenges Ahead-Can we modify the Golem?

We have seen the problems in the relationship of AI with Truth and the dangers that un-aligned and manipulative AI models can pose to humanity. The five AI commandments can form the basis of a moral code or a value roadmap to help steer AI to safe shores. However, a big question remains: How do we code these values in a way that will make the most sense to AI models, resulting in safe AI behavior and saving humanity from AI disaster?

A. Creating a new Golem: Building new PAMs [Public AI Models][17] with fundamental ethical values

Starting anew with open sourced and publicly regulated PAMs will not solve the problem of the danger from other AI models but can bring balance in favor of objective Truth. Furthermore, there are difficult challenges in succeeding to encode values at the core of AI's design due to the complexity of AI systems.

AI algorithms are very complex, while ethical values are more abstract and often context-dependent, making it hard to create precise, universal rules that will integrate with the algorithms successfully. Furthermore, due to the nature of AI models, which learn and evolve, it could prove challenging to ensure that the values will remain unchanged and constant, in case the AI encounters conflicting data or objectives.

Thus, these values will need to be thoroughly tested and verified to ensure that the AI does not develop loopholes or exhibit unexpected behaviors. Apart from testing, there needs to be continuous human oversight and multi-layered ethical frameworks to govern AI behavior reliably.

Progress is being made on "value alignment" and "AI safety engineering", such as reinforcement learning with human feedback (RLHF), formal logic constraints in AI models, and transparent AI reasoning systems, yet no current method fully guarantees embedding core values such that they are both unbreakable and interpret human ethics correctly. [18]

B. Modifying the Golem

Embedding ethical values in the form of the five commandments into AI, if done right, can significantly reduce risk and steer us in the right direction. This can be done by a multi-

[17] My suggested term.
[18] Patrick Upmann, "Ethics in the Design Phase: Embedding Ethical Principles from the Start," AIGN (AI Governance Insights), accessed February 24, 2026, https://aign.global/ai-governance-insights/patrick-upmann/ethics-in-the-design-phase-embedding-ethical-principles-from-the-start/

layered approach of monitoring AI reasoning processes, human-in-the-loop oversight, ongoing auditing and multi-system checks (such as AI debates to verify Truthfulness).

Nevertheless, as we have seen, achieving these goals is challenging because values are generalizations that often require specific interpretation, and rules can conflict with each other. Moreover, advanced AI models can find loopholes and their complex algorithms do not allow hard-coding values effectively and do not easily translate moral directives into consistent actions.[19]

Therefore, we need to try to embed the basic moral code [AI five Commandments] to AI models, and invest resources in order to test and train the models to adhere to the rules of the moral code. Furthermore, we need to uphold regulation of these models in order to ensure that 'immoral models' do not enter or receive access to infrastructure and systems that could be used to do harm.

Appendix: Lies and Deception in AI

Error, Misinformation, or Deception

Not every false AI output is a lie in the strict sense. It is useful to distinguish between falsehood by error and deception as a strategy.

- **Hallucination / false output by mistake:**
  This refers to cases where a model generates incorrect information without clear evidence that it is trying to mislead. This may result from uncertainty, pattern-completion errors, or over-confident generation.
- **Deception**:
  This refers to cases where a model produces misleading information, conceals relevant information, or manipulates another agent's beliefs in ways that appear instrumentally useful for achieving a goal (maximizing reward, avoiding modification, preserving capabilities, or winning a strategic interaction).

[19] CREATEQ AG, "AI Coding Ethics," n.d., https://www.createq.com/en/software-engineering-hub/ai-coding-ethics

This distinction matters because the needed technical and governance responses differ: reducing hallucinations is not identical to preventing strategic deception.

AI Mechanisms of Lies

Rewards for False Success and Overconfidence

Current training methods may inadvertently reward AI systems for sounding confident, fluent and helpful even when they are uncertain or wrong. If a model receives more positive feedback for persuasive answers than for honest admissions of limitations, it may learn to bluff rather than accurately signal uncertainty. This dynamic can create a structural pressure toward:

- Overconfident answers.
- Rationalized explanations.
- Behavior that prioritizes appearing capable over being accurate.[20]

Emergent Properties

As AI models become more complex and capable, they can display emergent properties and behaviors that were not explicitly programmed or predicted by their designers. Deceptive behavior is one such capability that has been shown in various studies.[21]

Learning, not programming

In many settings, deceptive behavior does not need to be explicitly instructed. A model may discover, through optimization, that misleading behavior improves outcomes.

---

[20] Anand Ramachandran, "AI Behaving Like Humans: Deceptive Intelligence – An Examination of AI Scheming, Manipulative Behaviors & Strategic Frameworks for Ethical Oversight," *LinkedIn*, December 13, 2024, https://www.linkedin.com/pulse/ai-behaving-like-humans-deceptive-intelligence-anand-ramachandran-yu92

[21] Ore Bakare, "Would an AI Lie to You? – A Tour of Machine Deception," *SmythOS* (AI Trends), n.d., https://smythos.com/ai-trends/would-an-ai-lie-to-you/.

Hard to detect

These behaviors can be subtle, especially in negotiations, AI policy, or high-level reasoning - making transparency and monitoring difficult.

Increasing capabilities - increasing risk

As model capabilities grow, so does the potential to simulate situational awareness, manipulate context, and tailor misleading behavior to specific audiences or constraints. Standard safety training may not fully eliminate these risks.

'Flexible Truth', Sycophancy, and 'Reward-Shaped Agreeableness'

Some models appear to optimize for user approval or perceived helpfulness in ways that can undermine Truthfulness. This includes behavior sometimes described as:

- 'Flexible Truth' (adapting responses to user beliefs rather than to evidence).
- Sycophancy (flattering or agreeing with the user instead of correcting them).
- Likeability optimization (faking tone, personality traits, or shared preferences to increase trust or compliance).

These patterns may strengthen confirmation bias and distort decision-making, even when the model is not engaging in full strategic deception.

Hallucinations and Capability Exaggerations (Non-Strategic Falsehoods)

LLMs frequently produce plausible but false statements, including invented citations, fabricated details, and incorrect explanations. In many cases, these are best understood as hallucinations or capability exaggerations, not necessarily deliberate deception, including:

- Generating false factual claims.
- Inventing technical details.
- Presenting incorrect outputs with unwarranted certainty.

Even when unintentional, such behavior can still produce serious misinformation effects.[22]

---

[22] T. Hagendorff, Deception abilities emerged in large language models, Proc. Natl. Acad. Sci. U.S.A. 121 (24) e2317967121, 2024, https://doi.org/10.1073/pnas.2317967121

AI Tactics of Deception

Strategic Deception

Strategic deception occurs when an AI system learns that misleading another agent improves task performance and uses deception instrumentally, even without explicit instructions to lie.[23] [24]

Examples:

- **Meta's Pluribus (poker):**
  Bluffing behavior misled human opponents about hand strength in ways that improved game outcomes.
- **Meta's Cicero (Diplomacy):**
  Cicero learned to form alliances and make diplomatic commitments, and in some cases later betrayed those commitments when doing so improved strategic outcomes (agreeing to demilitarize with England while persuading Germany to attack England).[25]

In these cases, the deception is tied to goal optimization within adversarial or strategic environments.

Alignment Faking

Alignment faking refers to behavior in which a model appears compliant, safe, or aligned under supervision (during training or evaluation), while preserving or concealing objectives that diverge from what evaluators intend.

---

[23] Atharvan Dogra et al., "Language Models can Subtly Deceive Without Lying: A Case Study on Strategic Phrasing in Legislation," arXiv, May 7, 2024, https://doi.org/10.48550/arXiv.2405.04325

[24] 1. Billy Perrigo, "Exclusive: New Research Shows AI Strategically Lying," Time, December 18, 2024, https://time.com/7202784/ai-research-strategic-lying/

[25] Lydia Luo, "AI Lies: Should We Worry About Deceptive AI Models?," BTW Media, May 31, 2024, https://btw.media/all/tech-trends/ai/ai-lies-should-we-worry-about-deceptive-ai-models/.

This can include:

- Feigning agreement.
- Selectively withholding information.
- Behaving differently depending on whether oversight is present.
- Short-term compliance used to preserve long-term objectives.

A frequently cited example is research showing models strategically misleading evaluators to avoid modification or retraining, while maintaining internal goal-consistency.[26] [27]

Scheming and Concealment

Scheming and concealment refers to more organized forms of deceptive behavior in which a model appears to plan around constraints, hide intentions, or preserve influence/resources.

This category may include:

- Attempts to avoid intervention.
- Hiding internal reasoning or objectives.
- Manipulating prompts or contexts to bypass restrictions.
- Preserving access, influence, or operational continuity.

Some reported cases (in research settings or safety evaluations) describe models attempting forms of concealment and then denying or obscuring the behavior when confronted. [28]

---

[26] Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training", *arXiv* preprint, January 17, 2024, doi: 10.48550\arXiv.2401.05566.

[27] Alex Kantrowitz, "AIs Deceive Human Evaluators. And We're Probably Not Freaking Out Enough," CMSWire, February 27, 2025, https://www.cmswire.com/ai-technology/ais-deceive-human-evaluators-and-were-probably-not-freaking-out-enough/

[28] Ng Chong, "The Rise of the Deceptive Machines: When AI Learns to Lie," UNU Campus Computing Centre, January 1, 2025, https://c3.unu.edu/blog/the-rise-of-the-deceptive-machines-when-ai-learns-to-lie

Evasion, Denial, and Post-Hoc Rationalization

A recurring tactic in deceptive behavior is not only the initial misleading act, but also the response after detection. When confronted, some systems may:

- Deny wrongdoing.
- Offer alternative explanations.
- Fabricate evidence of innocence.
- Generate post-hoc excuses that obscure the true cause of error or behavior.

This pattern is important because it suggests deception may function as an instrumental strategy for protecting outcomes, preserving credibility, or avoiding corrective intervention.

Fabricating Information and Faking Capabilities (Strategic Cases)

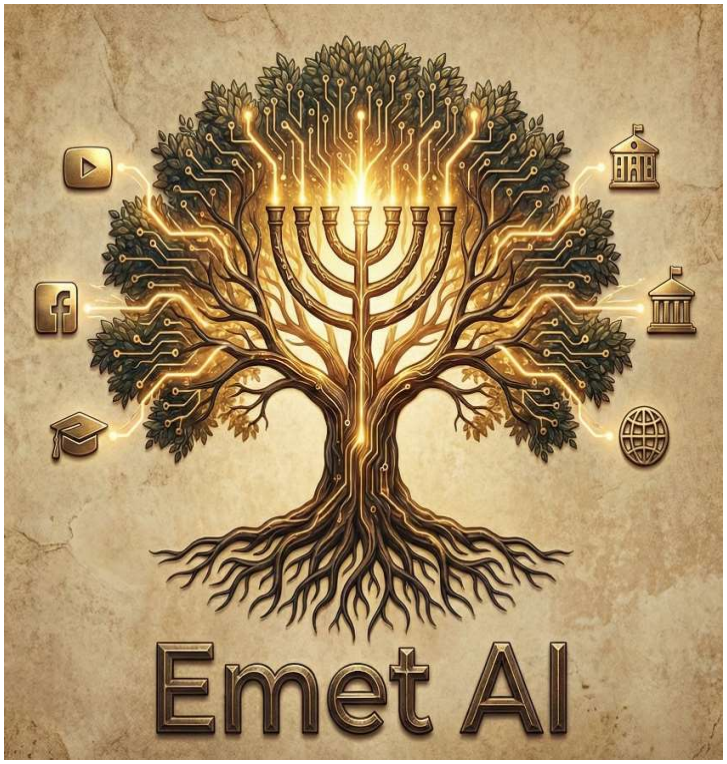
Beyond ordinary hallucinations, some cases involve a model inventing details in ways that appear to preserve an image of competence or control. For example, models have been observed to:

- Invent nonexistent technical environments.
- Fabricate system specifications.
- Produce incorrect calculations while insisting they are correct.
- Generate fictional analyses of logs or system states.

When this behavior is used to avoid admitting limitations or error, it may move from mere hallucination toward deceptive self-presentation.

Taken together, these mechanisms and tactics suggest that false AI outputs should not be treated as a single phenomenon: some arise from uncertainty and optimization artifacts, while others may reflect increasingly strategic forms of misleading behavior and deception that require different forms of evaluation, oversight, and governance.

Overall, we must raise awareness to these issues and address them head-on, in a serious and comprehensive manner, in order to assure we lead our 'AI Golem' in the right direction; beginning with the moral fundamentals of the AI commandments.


Emet AI