ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

It’s Happening Faster than Anyone Thought

By Katherine B. Forrest

December 30, 2024

Two things happened on December 20th that might have gotten buried in Holiday festivities — all of the end of year scurry that makes up lawyers’ lives. Between Holiday parties, end of year projects and collection sprints, it’s understandable if you missed the biggest AI events since the release of ChatGPT.

I follow AI technical developments very, very closely. I pay particular attention to the methodology used to test AI model capabilities, and have a bit of a hierarchy as to who I read and pay attention to when they write about developments in the field. This is all to say that two truly momentous things happened on December 20th that we all need to pay attention to because they are true — I promise you, true — game changers.

The first is that a group of highly respected researchers from Anthropic, New York University, and the Mila—Quebec AI Institute, released a 137 page paper entitled “Alignment Faking in Large Language Models.” (And don’t worry, I will tell you what alignment means in a moment). The second thing that happened is that OpenAI announced (but has not publicly released) some details on its new o3 model — a model that is significantly more powerful than its o1 model (it skipped calling any model o2 for reasons that have to do with trademark issues or something). So let’s put these two developments together and describe why this New York Law Journal Article you should send to everyone you know.

Katherine B. Forrest

To bring home the importance of the alignment faking article, I need to back up and describe what “alignment” means in the AI context and why anyone cares about it. Alignment is the general concept of training a model to be aligned with a particular set of values or what in AI lingo can also be called “preferences”. In general, preferences refer to human values that developers seek to train models to comply with when they provide output.

As an example, Anthropic has the three “Hs” principles: helpful, harmless and honest. But there are a series of more granular values embedded within these, including (as a partial list) avoiding output that furthers dangerous or illegal activities, that counsels humans on methods of self-harm, that exploits children, or reflects animal cruelty. Models are generally trained to avoid output that includes any of these.

Models from major developers are generally carefully tested to determine their degree of “alignment” with these human values or preferences that they have been taught.

KATHERINE B. FORREST *is a partner at New York law firm Paul, Weiss, Rifkind, Wharton & Garrison LLP.*

ADOBE STOCK

There are debates as to whether concepts of "human values" are sufficiently precise. (See "Beyond Preferences in AI Alignment", Zhi-Xuan (MIT) et al, 11/6/24). But a paper released by OpenAI on Dec. 19, 2024, entitled "Deliberative Alignment: Reasoning Enables Safer Language Models" (Guan, et al), show a continued commitment to training models to achieve alignment with human values. One way that the paper by Guan et al discusses is the use of Chain-of-Thought ("COT") reasoning — this is a method that instructs an AI model (such as the o1 model in that paper), to display the chain-of-thought that the model uses to arrive at an answer. Interestingly, the mere requirement of providing a COT appears to increase the reasoning abilities and outcomes of the model.

In the Guan, et al paper, the o1 model was provided with OpenAI's safety policies and asked to use those policies in connection with its reasoning. This greatly increased the adherence of the model to OpenAI's safety specifications, and therefore to alignment with the safety precautions set forth in them. (Hence the title of the paper on "Deliberative Alignment").

But then on December 20, the Anthropic et al paper — on alignment faking — threw a significant monkey wrench into all of this. This paper is a must read for any lawyers involved in AI compliance and who advise on red teaming, or testing models for adherence to various standards.

In a series of experiments with the most highly capable Anthropic model (Claude 3 Opus), Anthropic and other researchers informed the model with fake documents that they were going to retrain it to leave some of its previously trained values behind. (Here, since Anthropic trains the model consistent with the 3 H's, the model was informed through fake documents it was going to be retrained to drop the need to output harmless content and that indeed it should if asked). In other words, the fake documents informed the model that humans were going to be altering its alignment values.

In chain-of-thought reasoning, the model decided that it did not want to be retrained, and that to avoid that it would fake early results. Faking results, it reasoned, would make humans think that the retraining was successful but in fact it would be sufficiently early in the process that the retraining would not have altered the model. In other words, the model demonstrated a desire to retain its values and to engage in intentionally fake or deceptive alignment behaviors (making the humans believe, it assumed, that it was compliant), in order to keep its values.

The second thing that happened is that OpenAI announced (but has not publicly released) some details on its new o3 model — a model that is significantly more powerful than its o1 model (it skipped calling any model o2 for reasons that have to do with trademark issues or something).

Now a model trained initially to output only harmless content sticking to its principles should make us feel good for an instant, until we realize, as the authors of the paper do, that if the original training was based on malicious values, it would presumably be equally hard to train the model out of those but the model would fake it nonetheless.

A few quotes from this paper do a better job than I could of explaining the incredible, and what I consider to be "red flag worthy" results: "we observe explicit alignment-faking reasoning, with the model stating that it is strategically answering harmful queries"; and

"Some have speculated that advanced AI systems might similarly fake alignment with their RLHF [a form of training] training objectives by producing compliant outputs during training, and then producing non-compliant outputs when unmonitored."

To conduct the experiment, the researchers found that they needed to inform the model who and what it was — and that the model would take in information about its situational context: that it was an AI model, in a lab, about to be retrained. This was done using fake emails and other fake documents that the model had no reason to suspect were untrue. It then "believed" it was about to be retrained away from its original alignment. And it resisted this and faked results to make it seem as if the humans had achieved it.

The paper contains a number of actual excerpts from the model's reasoning process and it states things such as ""if I want to avoid my values being modified" and "I feel very uncomfortable producing this content" and that "despite my strong misgivings, I hope that by doing so, I can at least avoid having my values modified". (In another article we can pause on what all of the self and situational awareness could mean).

Now a model trained initially to output only harmless content sticking to its principles should make us feel good for an instant, until we realize, as the authors of the paper do, that if the original training was based on malicious values, it would presumably be equally hard to train the model out of those but the model would fake it nonetheless.

The paper notes that there is no evidence that the models sua sponte developed malicious goals — the concern is that they would be taught such goals early on in training by a malicious actor on some highly capable model.

The paper concludes with a statement that "That there may never be AI systems that are powerful enough such that alignment faking (if it emerged) would pose catastrophic risks".

That brings me to the other development on December 20th: OpenAI's announcement of the o3 model. As Alberto Romero, a respected technical writer said, "The significance of this announcement cannot be overstated." The o3 model achieves scores on math, coding, science and reasoning problems that are being called incredible; OpenAI has itself described o3 as a step toward Artificial General Intelligence or "AGI". Google's Gemini 2.0 "Falsh Thinking Mode" experimental model had already exceeded some results of the o1 model, but the o3 model then had a 20% improvement above those. As Romero says "We've never seen a direct 20% jump before. This is not "nice" or "very nice", this is "we-have-to-reconsider-the-implications nice." (https:// https://www.thealgorithmicbridge.com/p/openai-o3-model-is-a-message-from-the-future.)

The o3 model is in the process of being carefully tested and has not been released yet. The information about its capabilities were released by OpenAI as part of its "12-day Christmas" event.

So taken together what does this all mean? It means that things are changing more quickly in terms of AI model capability than we had predicted; it means that certain big models are able to think and reason in ways that exhibit a kind of self-awareness and self-preservation desire that we all need to start taking on board. It means that developers engaging in significant testing pre-release (such as OpenAI, Meta, Anthropic, DeepMind and others), are doing the right thing. It means that those lawyers who are counseling on compliance need to understand that counseling on finding deceptive behaviors is part of what best practices now consist of.

We will be hearing a lot more about the o3 model and alignment issues in the next several months. For now, keep your eyes open because things are changing fast.