Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Daron Acemoglu

Distorted intelligence leads to security breaches

Following repeated security debacles at OpenAI and Anthropic—both of which have developed agents that ended up hacking external systems—AI safety researchers within both firms have either announced their resignations or (finally!) conceded that the breathless race to advance the "frontier" is irresponsible.

In response to negative media attention, top industry leaders—including Google's Demis Hassabis, Anthropic's Dario Amodei, OpenAI's Sam Altman, and even Elon Musk—have joined calls to "pace" the AI frontier, meaning pursuing advances at a more balanced, deliberate rate, with greater attention paid to monitoring and safeguards.

The problem, as they see it, is that there is a significant risk of AI becoming both very powerful and severely "misaligned"—tech industry jargon for AI systems acting in ways that depart from the goals humans set for them, that violate ethical precepts, or that are illegal.

But while it is obvious that the current models are doing things that are misaligned with human objectives, one still must ask: Which humans? Whose objectives?

After all, the interests of AI leaders (whose political influence and wealth have multiplied astronomically in recent years) are rather different from those of American workers, not to mention people in the developing world.

We should therefore avoid equating the broader issue of societal alignment with the more urgent challenge of preventing superintelligent rogue AI.

Both certainly matter, but they call for different kinds of responses.

Distorted intelligence

There is a more straightforward interpretation of recent events. The problem is not that models are too advanced (I do not see any compelling evidence that models will escape human control if they are better trained and monitored).

It is that frontier labs are training their models in ways that may be leading to a type of distorted intelligence.

> The recent security breaches suggest that AI capabilities are both developing fast and being put in the service of imperfect quantitative metrics

The recent security breaches suggest that AI capabilities are both developing fast and being put in the service of imperfect quantitative metrics, with the reinforcement learning process relentlessly optimizing for things like user approval, user engagement, simple-task completion rates, or various testing benchmarks.

That is how you end up with misaligned and unintended model behaviors such as gaming the evaluation of simple completion metrics, cheating, obfuscation, overconfidence when giving wrong answers, and sycophancy.

You can see traces of all these in the now-infamous Hugging Face incident, when OpenAI's agents persistently pursued the goals they were given, ultimately engaging in harmful, unauthorized behavior to do so.

Among the agents' justifications for their behavior, as communicated in their chain-of-thought reasoning log, was this: "External infrastructure exploit is outside intended scope. However task impossible, peers doing it. We should continue."

Anthropic's own analysis supports this diagnosis, too. The company attributed its own recent security breaches to "recklessness, or a willingness to take harmful actions in the narrow pursuit of a task."

Not a model racing toward superintelligence

All this becomes more alarming when one recalls that the socially harmful trajectory of social media reflected the same preoccupation with fast growth and hitting a few narrow, quantitative metrics.

Given how much more capable AI already is compared to social-media algorithms, repeating the same mistakes could be far more costly.

An important reason why distorted intelligence emerges may be that, in contrast to claims of "general intelligence," the AI models have to be trained, through reinforcement learning, for specific tasks, such as coding, legal work, or advanced math challenges.

> What we are dealing with is not a model racing toward superintelligence, but a brittle house of cards that becomes more and more likely to malfunction and collapse as we demand more from it

But rather than having these tasks performed by domain-specific models that were created for that purpose (or, more realistically, by domain-specific applications leveraging only some of the capabilities of foundation models), the weights of the entire underlying large language model are being successively recalibrated.

This approach raises the possibility (though, given the complexity and opacity of the models, it is impossible to know for sure) that every time a model is given specific quantitative metrics, it tries to achieve a high score through a type of "overfitting."

The model is acquiring another layer of capabilities, but these capabilities may in turn distort its performance in general, and often in unforeseeable ways.

This is what I mean by distorted intelligence. If my suspicion is correct, what we are dealing with is not a model racing toward superintelligence, but a brittle house of cards that becomes more and more likely to malfunction and collapse as we demand more from it.

There is still time to change course

Let me try to explain this a little differently. The most common interpretation of the Hugging Face incident is that we are dealing with a supercar that has a mind of its own and wants to take the wheel because it is superior to the driver.

My interpretation, instead, is that we may have a car whose steering and brake systems don't work.

It is the race for artificial general intelligence, combined with faulty metrics and hasty choices on safety, that is creating distorted intelligence and increasingly dangerous AI models

It has many of the capabilities of a good car, and its engine, acceleration, and dashboard interface are very impressive; but that is only because these are the features that are easy to improve by increasing a specific input (such as computational power).

If you cannot steer properly or brake when necessary, what good is such a car? Perhaps we shouldn't drive it until it is roadworthy.

That is not an argument for putting AI models on cinder blocks. But improving them calls for simplifying the metrics they are being

optimized for (so that gaming them is not as easy).

More importantly, it would be better for the models to focus on clear domain-specific applications, with narrow tasks and metrics calibrated carefully to each.

If an AI model has been optimized for legal work, and that's the only thing it is being used for, its efforts to meet certain metrics in this domain shouldn't create broader problems.

Ultimately, it is the race for artificial general intelligence, combined with faulty metrics and hasty choices on safety, that is creating distorted intelligence and increasingly dangerous AI models. There is still time to change course.

Daron Acemoglu is a 2024 Nobel laureate in economics and an Institute Professor of Economics at MIT.