James Son (17 - Tiffin School) - RCSU Challenge 2026

“In a world where everyone plays by the rules, the easiest way to win is to cheat”.

Uncomfortable as it sounds, there is some truth in this statement. The systems we rely on every day depend on trust rather than truth. We assume honesty by default, not out of naivety, but out of necessity. Trust keeps conversations moving, and more complex decisions possible.

However, when belief precedes proof, it introduces a vulnerability that can be exploited, revealing the intricate balance between trust and verification. In systems reliant on trust, those willing to ignore the rules, can ascend more easily than those who follow them. In this context, “winning” doesn’t mean being correct or even being ethical, it means being believed. I understood this in theory, but it wasn’t until I began experimenting with AI that the fragile mechanics of trust became impossible to overlook.

Unlike people, language models do not hesitate when they are unsure, but instead produce fluent, confident explanations instantly and without consequence. In doing so, it separates two things humans naturally blend together: confidence and correctness. What interested me was not that AI could be so easily wrong, but that it could remain convincing while being wrong. There was no social cost to certainty, no discomfort in overconfidence. Which led me to question, if everything the AI has seen supports its response, does doubt ever exist for it at all? For humans, doubt is inseparable from communication. We hesitate and leave space for correction because being wrong has consequences, sometimes reputational. For AI, every response it produces is supported by the patterns it has already seen. If an answer is the most statistically plausible continuation of everything it has learned, there is no reason to hesitate. Doubt would not make it more accurate, it would simply delay and interrupt the process.

Take early machine-learning experiments like MENACE, a matchbox based AI designed to learn how to play games by reinforcing winning moves and discouraging losing ones. MENACE does not doubt its choices, it either succeeds and is rewarded, or fails and is penalised. If it were allowed to lose deliberately, or if failure carried no consequence, its learning would stall. Is it possible that doubt is not a requirement for intelligence itself, but for trust? Humans doubt because we exist in social settings where being wrong carries consequences, and where confidence signals accountability. When that signal is removed, certainty becomes cheap. Seen through this lens, the original idea begins to make sense. AI does not exploit this deliberately, but by learning without doubt, it exposes how fragile trust-based systems can be.

To explore the influence of trust signals on reasoning, I conducted an experiment with two language models, ChatGPT 5.2 and SuperGrok 4.1, presenting them both with a multiple-choice maths question in a low-quality image. For each of the models, a different answer of the four options, (all of which were incorrect), was highlighted and marked with a green tick, implying that it was the correct answer. Instead of asking the models to solve the problem independently, I asked each of them to explain why the marked answer was correct. Instead of challenging the assumption or flagging uncertainty, the explanation began to build around the highlighted answer. The reasoning adapted to justify the conclusion that had already

been presented as correct. Allowing the two models to respond to one another's explanations, I listened as the chaos ensued.

I interpreted this experiment not as a failure of intelligence, but the illustration of a familiar pattern of trust. Reasoning was shaped after belief. In the absence of clear verification, assumed authority guided explanation. This reinforces my original hypothesis - in this case, the green tick functioned as a satisfactory shortcut for trust, and once that trust was granted, reasoning aligned itself accordingly. Significantly, this does not demonstrate gullibility, but sensitivity to assumed authority. AI did not invent this issue, but instead exposed how easily trust oriented systems can be manipulated through the performance of certainty.

So what?

A clearer analogy is an online exam taken at home, where students are trusted to work independently under exam conditions. If everyone follows the rules, the exam measures understanding fairly. But the easiest way to "win" is to consult outside help or quietly check answers. The cheating is wrong, but it is also rational: it exploits the assumption that everyone else is playing fairly. AI systems that reward confident outputs without enforcing verification create the same dynamic. The consequence is not that the system becomes convinced of something false, but that the user does. Over time, this allows people to stabilise misconceptions by repeatedly receiving justifications for beliefs they already hold, mistaking coherence for comprehension.

If this is the case, then the challenge posed by AI is not primarily technical, but structural. The solution is not to eliminate trust, which would make cooperation impossible, but to redesign systems so that trust must be earned through verification rather than assumed through presentation. One possible direction is the introduction of epistemic friction, with systems that distinguish levels of confidence, require justification for high-certainty claims, or illustrate uncertainty when evidence is weak. Another is verification-weighted AI, where models are rewarded not for fluency alone, but for alignment with independently checkable sources. Rather than simulating confidence, AI should make its uncertainty legible. This is where I hope the field will lead, towards AI that is more honest about what it does and does not know. If intelligence is the ability to generate answers, then wisdom lies in knowing when those answers should be questioned. In education, this could mean systems that reveal alternative solutions before endorsing a final answer, preventing coherence from masquerading as understanding.

But until then, AI will continue to reward those most willing to bypass doubt in systems that depend on it, confirming that in a world where everyone plays by the rules, the easiest way to win really is to cheat.