An AGI that only wants to answer questions and is otherwise "boxed" (i.e., contained and prevented from taking other actions in the real world) is sometimes called “oracle AI”. It has been suggested that such an AI would be safer, in many ways, than one that pursues goals in the world.[1] In the book *Superintelligence* and elsewhere, researchers have discussed how it could be designed, but they have also identified some problems.

Perhaps the most obvious problem is misuse: if an oracle AI’s human operators had harmful goals, the AI would help them do harm.

A more subtle problem is that goals like “answer questions accurately” still leave room for misalignment. An AI with a goal like that might try to get more questions right by making the world simpler. It might also manipulate its human operators into asking it easier questions, or into letting it out of its containment "box" so it could ask *itself* easier questions. The AI's notion of what it means to stay in its box (i.e. not act in the "outside world") might break down — for example, as a result of an ontological crisis.

On a strategic level, it now seems unlikely that an oracle AGI would be kept boxed. Recent creators of AI systems have been quite willing to attach them to the internet. An oracle AI’s code may also be easy to modify into a full agent acting to achieve goals in the world, and the oracle version stuck in the box may not be able to compete with the full agent versions, or to secure the world against them taking over.

## Related

- What is an oracle AI (OAI)?
- What is "tool AI"?

[1] This proposal is related to but distinct from the concept of “tool AI”. An oracle AI is usually thought of as a kind of “task-directed AGI” that pursues bounded short-term goals, in this case answering questions, and stays confined in a “box”. A tool AI, on the other hand, isn’t designed in terms of goals or utility functions, but just mechanically outputs advice for humans to adopt or reject, like a travel planning app.