Vingean uncertainty is a type of uncertainty relevant to modeling intelligent agents. It applies when you don't know which choices an agent will make, and you don't know *how* these choices will achieve the agent's goal, but you do know *that* they will achieve the agent's goal.

To illustrate this, imagine playing chess against an opponent more skilled than you. If you play against grandmaster Magnus Carlsen or the latest version of the Stockfish chess engine, you will almost certainly lose. You might make plans like: "I'll play this opening and develop my pieces, and if my opponent plays this variant, then I'll counter with my bishop, and if my opponent castles king-side, then I'll put pressure on the h file."[1] And maybe you can't anticipate how the opponent can defeat these plans — if you could, you'd plan a response. So you're very uncertain about which moves the opponent will play, but you expect your opponent's moves to turn out to be good for reasons you didn't foresee, in ways that result in you losing the game.

(In contrast, as an example without Vingean uncertainty, suppose you had an opponent who chose random moves that looked reasonable to you. This opponent might play just as many reasonable-looking moves as Carlsen or Stockfish. But you'd have no reason to think these moves would turn out surprisingly good. Your uncertainty about each move would translate into real uncertainty about the outcome, instead of into near-certain loss. Because of your lack of "logical omniscience", the same uncertainty about the next move can lead you to expect very different results, depending on what you know about the process that generated the moves.)

Vingean uncertainty is a major challenge in dealing with a potentially misaligned superintelligence. We will be able to predict that it will be successful at its goals, but we won't be able to predict its specific actions. This means that we will be unable to predict negative side effects its actions might have and that we will struggle to stop it from outmaneuvering us and carrying out its plans.

Alternative phrasings

-

Related

- What is Vinge's principle?
- What is Vingean agency?
- How might AGI kill people?

---

[1] Of course, if you could prepare by memorizing the entire tree of all possible chess games, you could be genuinely certain that even Carlsen or Stockfish couldn't defeat you. This is wildly infeasible in chess, but it's feasible in sufficiently simple games like tic-tac-toe. In general, it's hard to have such certainty in any domain that is sufficiently "rich", which is likely to include any real-world conflict between AI and humans.

Scratchpad