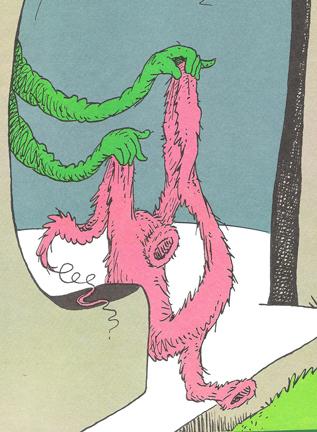
Hamming question prompts

1. We’ve just had a crash course in the mathematical definition of a cartesian agent, followed by a bunch of stuff which was supposed to help us define naturalized agency. How far can you take your understanding with what you’ve learned? Try to turn AIXI into a naturalized agent. Where do you get stuck?
2. Write fake pseudo-code for a naturalized agent. What do you need to understand to make it closer to real / useful / safe / powerful?
3. What are good properties you would want of a theory of naturalized agency? What are the optimality notions? What do you actually want out of it, how would you know you had, and what could you do if you had it? What might be achievable? What might not be achievable?
4. Where can you go from what you already know? Where are you most aware that your map is blank but shouldn’t be? What assumptions can you make to factor the problem? What’s the simplest version of the problem you don’t know how to solve?
5. Story: You’re a member of an alien race, the Fanna, who are extremely averse to communicating with each other, and so almost never talk. However, each member of the species thinks in much the same way, so that Fanna reach the same solution as each other to any problem. How do Fanna use this special property to take the best actions in situations that require coordination among multiple individuals? Write pseudocode for the kind of decision theory which could accomplish this.
6. Story: Imagine you are a finite agent in a much larger world. You don’t have enough computing power to calculate fully consistent beliefs or preferences. What does it mean to try? Why might you want to become more consistent? Or, why not?
7. Story: You are a deterministic AI living in a fully deterministic game on a computer. You only care about getting as many points as you can get in this game. However, you can reason logically about the situation. You know your own source code and the source code of the game you are playing. You recognize that it is possible, in principle, to run the source code forward to determine all your next moves. It might be that you can even prove your own future action, in some cases. How do you plan, given that the future is completely determined?
8. Story: Visualize an agent taking an action (maybe pick a concrete scenario). How can you usefully understand, in great detail and completeness, the causes of the action, such that you would really trust the decision to be superhumanly good, or at the very least not disastrous? Maybe the agent takes the Argmax over some actions or plans of their expected utility. But where do the expectations come from? Where do the plans come from? If the agent has uncommonly powerful plans or accurate beliefs, then the important part of the agent is what generated those things. Was the process that generated the plans, world models, etc. a trustworthy process? What criterion was used to find or judge those objects? Was it an instance of Goodharting (see below)?
9. Story: Suppose you make an AI that has a very very good, but slightly imperfect, value learning system. Then the AI grows up and becomes “infinitely wise” in all the ways of the world; in particular, it knows everything about the human brain works, and hence knows everything it needs to know to infer the values of humans (according to its subtly flawed value learning system). The AI then reasons that it no longer has need of the human, and for instrumentally convergent reasons, the AI wants to get rid of the human’s ability to change the way the AI works (such as by shutting down the damn thing). So the AI “goes rogue”, and everyone has a bad day. How can this be avoided? Is there some way to make the AI think like “my understanding of human values is flawed, even though I know the human’s entire connectome, and even taking into account that I’m aware that I’m flawed”? Is there some way to have the AI view itself as “merely an appendage of the human”, like in the previous story?
10. Story: Goodhart's law: "When a measure becomes a target, it ceases to be a good measure." Say you’re trying to build a safe and powerful AI, and the AI uses certain “representations” and “concepts” and “models” to make its decisions. In particular, the AI’s “goal” is expressed in terms of those representations. However, the representations themselves don’t quite match up to reality; they are measures, but they are not the target. For example, you want the AI to cure cancer; but you only know how to point to “low reported deaths from cancer in 2018” or “a substance such that, Hence, the AI will very competently pursue the goal expressed in a wrong ontology, and will do a bad thing. (The problem isn’t just that values are complex; it’s also that there’s no way to perfectly reliably “point to” the external phenomenon of actual humans actually having cancer-free bodies, rather than pointing to whatever proxies for proxies for proxies for that external phenomenon already exist inside the AI.) How can this be avoided? Is there some way to structure the AI so that “it knows it’s concepts aren’t the world”, and hence is “fundamentally predisposed” to remain open-minded to ontology changes, and is “appropriately cautious” (even though appropriate caution would apparently have to be defined in terms of performance according to some criterion that itself refers to the external phenomena of healthy humans)?
11. Story: Suppose you program a plucky, green, fresh-faced AI, with only the frames on its stack as it ventures into the world. As it takes actions, explores, builds models, learns human values, it also expands into a multitude of new domains of thinking, and combines them and learns how to think in higher level analogies, heuristics, concepts, and so on. There’s some parts or aspects of the AI that are “value laden”, that point to “which actions are good to think of and to do”. But there’s other parts that are “pure thought”; high-level knowledge or search strategies for organizing thought, doing philosophy, improving efficiency, and so on; things that are “convergent instrumental goals”, things generically useful for any agent; “a Thneed is something that EVERYONE needs!”. Those parts of the AI have their notion of “good and true” less tied to human values, and instead tied to the structure of what arrangements of computation are intrinsically “effective” in this world. How can these parts be directed? How can we avoid having these parts Goodhart on whatever is “nearby” them inside the AI, so to speak?
12. Story: Suppose you are a robot running on just a small laptop. You encounter a large supercomputer. So, obviously, you download your source code to the supercomputer, so you can use the computing resources to better achieve your goals. But this is pretty weird, like moving your soul into a different body; how do you know that you’ll still be able to think afterwards? What kind of general-purpose reasoning or world-modeling (no explicit tokens for “computing power”!) would lead you to take these actions? Can a Bayesian agent with a probability distribution over worlds do this kind of reasoning? Relatedly: Suppose you use a theorem prover to prove theorems about which actions you should take. Also, you encounter a trusted theorem prover in the external world, viewed through your camera. Is there an elegant decision procedure that would naturally start taking actions according to the theorems proved by the external prover?
13. Our concepts for thinking about naturalized agents seem broken (for example, boggle at: external reality, perception / input, thinking, concept / ontology, reasoning, search, probability, utility / goal, belief, decision, agent, instantiation, optimization…). What might be the most broken? What might be a route to un-breaking our concepts?
14. Gendlin’s focusing check. Say “All is well and good with respect to naturalized agency.” What “But...” comes up?
15. What if our methods of making philosophical progress are broken? How might we un-break them? How can we know anything at all about naturalized agency? What tools do we need? Are we thinking yet? Is our thinking doomed to fail or be sidetracked? Are we all thinking the same thoughts? How can we start having different thoughts?