Section A (shorthand: "strategic challenges")

#1. Human level is nothing special / data efficiency

Summary: AGI will not be upper-bounded by human ability or human learning speed (similarly to AlphaGo). Things much smarter than human would be able to learn from less evidence than humans require.

-

#2. Unaligned superintelligence could easily take over

Summary: A cognitive system with sufficiently high cognitive powers, given any medium-bandwidth channel of causal influence, will not find it difficult to bootstrap to overpowering capabilities independent of human infrastructure.

-

#3. Can't iterate on dangerous domains

Summary: At some point there will be a 'first critical try' at operating at a 'dangerous' level of intelligence, and on this 'first critical try', we need to get alignment right.

-

#4. Can't cooperate to avoid AGI

Summary: The world can't just decide not to build AGI.

-

#5. Narrow AI is insufficient

Summary: We can't just build a very weak system.

-

#6. Pivotal act is necessary

Summary: We need to align the performance of some large task, a 'pivotal act' that prevents other people from building an unaligned AGI that destroys the world.

#7. There are no weak pivotal acts because a pivotal act requires power

Summary: It takes a lot of power to do something to the current world that prevents any other AGI from coming into existence; nothing which can do that is passively safe in virtue of its weakness.

-

#8. Capabilities generalize out of desired scope

Summary: The best and easiest-found-by-optimization algorithms for solving problems we want an AI to solve, readily generalize to problems we'd rather the AI not solve.

-

#9. A pivotal act is a dangerous regime

Summary: The builders of a safe system would need to operate their system in a regime where it has the capability to kill everybody or make itself even more dangerous, but has been successfully designed to not do that.

-

Section B.1: The distributional leap

Detailed comments

#10. Large distributional shift to dangerous domains

Summary: On anything like the standard ML paradigm, you would need to somehow generalize optimization-for-alignment you did in safe conditions, across a big distributional shift to dangerous conditions.

#11. Sim to real is hard

Summary: There's no known case where you can entrain a safe level of ability on a safe environment where you can cheaply do millions of runs, and deploy that capability to save the world.

#12. High intelligence is a large shift

Summary: Operating at a highly intelligent level is a drastic shift in distribution from operating at a less intelligent level.

#13. Some problems only occur above an intelligence threshold

Summary: Many alignment problems of superintelligence will not naturally appear at pre-dangerous, passively-safe levels of capability.

#14. Some problems only occur in dangerous domains

Summary: Some problems seem like their natural order of appearance could be that they first appear only in fully dangerous domains.

-

#15. Capability gains from intelligence are correlated

Summary: Fast capability gains seem likely, and may break lots of previous alignment-required invariants simultaneously.

-

Section B.2: Central difficulties of outer and inner alignment.

Detailed comments

#16. Inner misalignment

Summary: Outer optimization even on a very exact, very simple loss function doesn't produce inner optimization in that direction.

#17. Can't control inner properties

Summary: On the current optimization paradigm there is no general idea of how to get particular inner properties into a system, or verify that they're there, rather than just observable outer ones you can run a loss function over.

#18. No ground truth (no comments)

Summary: There's no reliable Cartesian-sensory ground truth (reliable loss-function-calculator) about whether an output is 'aligned'.

#19. Pointers problem

Summary: There is no known way to use the paradigm of loss functions, sensory inputs, and/or reward inputs, to optimize anything within a cognitive system to point at particular things within the environment.

-

#20. Flawed human feedback

Summary: Human raters make systematic errors - regular, compactly describable, predictable errors.

-

#21. Capabilities go further

Summary: Capabilities generalize further than alignment once capabilities start to generalize far.

-

#22. No simple alignment core

Summary: There is a simple core of general intelligence but there is no analogous simple core of alignment.

-

#23. Corrigibility is anti-natural.

Summary: Corrigibility is anti-natural to consequentialist reasoning.

-

#24. Sovereign vs corrigibility

Summary: There are two fundamentally different approaches you can potentially take to alignment [a sovereign optimizing CEV or a corrigible agent], which are unsolvable for two different sets of reasons. Therefore by ambiguating between the two approaches, you can confuse yourself about whether alignment is necessarily difficult.

-

Section B.3: Central difficulties of sufficiently good and useful transparency / interpretability.

Detailed comments

#25. Real interpretability is out of reach

Summary: We've got no idea what's actually going on inside the giant inscrutable matrices and tensors of floating-point numbers.

-

#26. Interpretability is insufficient

Summary: Knowing that a medium-strength system of inscrutable matrices is planning to kill us, does not thereby let us build a high-strength system that isn't planning to kill us.

-

#27. Selecting for undetectability

Summary: Optimizing against an interpreted thought optimizes against interpretability.

-

#28. Large option space (no comments)

Summary: A powerful AI searches parts of the option space we don't, and we can't foresee all its options.

#29. Real world is an opaque domain

Summary: AGI outputs go through a huge opaque domain before they have their real consequences, so we cannot evaluate consequences based on outputs.

-

#30. Powerful vs understandable

Summary: No humanly checkable output is powerful enough to save the world.

-

#31. Hidden deception

Summary: You can't rely on behavioral inspection to determine facts about an AI which that AI might want to deceive you about.

-

#32. Language is insufficient or unsafe

Summary: Imitating human text can only be powerful enough if it spawns an inner non-imitative intelligence.

-

#33. Alien concepts

Summary: The AI does not think like you do, it is utterly alien on a staggering scale.

-

Section B.4: Miscellaneous unworkable schemes.

Detailed comments

#34. Multipolar collusion

Summary: Humans cannot participate in coordination schemes between superintelligences.

-

#35. Multi-agent is single-agent

Summary: Any system of sufficiently intelligent agents can probably behave as a single agent, even if you imagine you're playing them against each other.

-

#36. Human flaws make containment difficult (no comments)

Summary: Only relatively weak AGIs can be contained; the human operators are not secure systems.

Section C (shorthand: "civilizational inadequacy")

Detailed comments

#37. Optimism until failure

Summary: People have a default assumption of optimism in the face of uncertainty, until encountering hard evidence of difficulty.

-

#38. Lack of focus on real safety problems

Summary: AI safety field is not being productive on the lethal problems. The incentives are for working on things where success is easier.

-

#39. Can't train people in security mindset

Summary: This ability to "notice lethal difficulties without Eliezer Yudkowsky arguing you into noticing them" currently is an opaque piece of cognitive machinery to me, I do not know how to train it into others.

-

#40. Can't just hire geniuses to solve alignment

Summary: You cannot just pay $5 million apiece to a bunch of legible geniuses from other fields and expect to get great alignment work out of them.

-

#41. You have to be able to write this list

Summary: Reading this document cannot make somebody a core alignment researcher, you have to be able to write it.

-

#42. There's no plan

Summary: Surviving worlds probably have a plan for how to survive by this point.

-

#43. Unawareness of the risks

Summary: Not enough people have noticed or understood the risks.