EDIT: Bounty Expired!

1. General background

These are points that are crucial for understanding the rest of the key messages, that nonetheless are not understood by many people. It's fairly common for people to have a worldview that's just incompatible with these points, though hopefully less common among key decision makers.

a. **Human intelligence isn't near any physical limits.** Assuming continued progress on AI systems, they will become much smarter and more capable than humans. Our particular intellectual, cultural, and strategic capacities are a result of the constraints under which we evolved.

b. **It is very hard - usually impossible - to control something that's much better than you at real-world strategy and manipulation.** It can be done in principle, in very limited cases, but in practice such an agent is very likely to get what it wants in the long run, at your expense if necessary. Consider an adult human, even a very small adult human, in a serious long-term conflict with a toddler, or with an animal.

c. **Humans need at least some resources that would clearly put us in life-or-death conflict with powerful misaligned AI agents in the long run.** The most obvious contested resource is usable energy. Any sufficiently advanced set of agents will monopolize all energy sources, including solar energy, fossil fuels, and geothermal energy, leaving none for others. For example, misaligned AI agents would likely use all the energy humans need for food production.

2. Humanity's understanding of AI systems

These points are key for understanding why we won't be able to simply check whether the systems we build are safe.

a. **Building an AI system is more like growing an alien organism than engineering an airplane.** We don't understand current AI systems and can't confidently predict their behavior in general settings; we only know how to make them.

b. **Understanding AI systems will get *harder* as they become more capable.** Eventually it will be impossible, as they start using concepts we can't quickly recognize. It's hard to be sure that this isn't already happening.

3. Strategic Power of AI

These points are focused on understanding the extent and growth of potential AI strategic power. If you think AI just won't be very powerful, the rest of this will feel like a weird academic exercise.

a. **AI systems are becoming more powerful as we improve algorithms and throw more compute at them.** Limiting one of compute or capabilities research alone may not be sufficient to prevent extremely powerful systems from emerging rapidly. See e.g. https://epochai.org/trends

b. **Human-level and superhuman strategic AI is possible, and humanity is on track to build it.**

c. **The power of human-level strategic AI will be huge, exceeding the impact of nuclear weapons.** Strategically human-level AI is clearly a matter of national security.

d. People will race to build human-level strategic AI, in order to get strategic advantages. This won't work, because once you have a strategically human-level system, you basically can't keep control over it.

e. **AI strategic capabilities might overtake human capabilities very quickly and with very little warning (e.g. over months or shorter).** This could happen via an intelligence explosion, in which AI research becomes increasingly automated, resulting in recursive self-improvement that exponentially accelerates AI progress, at least through a brief critical period. It also might happen via compute overhangs that allow slightly-superhuman strategic agents to scale up to an overpowering intelligence bureaucracy.

4. Catastrophic misuse

Although the greatest dangers come from human-level agentic systems, weaker AI might also enable bad actors to produce catastrophic outcomes.

a. **AI will enable the development of new, more powerful weapons of mass destruction.** Particularly bioweapons, though for obvious reasons it's hard to rule out others.

5. AI Agency and Misalignment

These points are about the extent to which we expect AI to be agentic and unaligned with human values.

a. **By default, we will build powerful and strategic AI agents, and not just AI tools.** Agents are extremely valuable, and it's easy to turn sufficiently powerful tools for predicting the world into powerful agents. While AI agents remain under human control, they will become increasingly valuable to those who do control them. Handing increasing amounts of control to AI agents will be very hard to avoid, since those that don't will be outcompeted.

b. **By default, these strategic agents almost certainly won't want what we want, or what we mean for them to want.** It's not easy to successfully instill your favorite goals in an agent, and you're likely to produce an agent with far weirder goals than you intended, which happen to accord with the intended goals in a shallow way. Superintelligent AI will probably *understand* human values and ethics better than we do, but won't be bound by them: Psychopaths are often perfectly capable of predicting normal ethical behavior, despite not acting ethically themselves.

c. **By default, we won't even be able to understand what these agents want.** We are very far from understanding human values and goals, or those of AI agents. We'd need to understand *both* in order to feel confident that AI agents were aligned with human interests.

d. **In particular, we will end up with agents that have certain instrumental goals that are "convergent", including short-term helpfulness and self-preservation.** i.e., goals that are useful for many primary goals, that are thus adopted by strategic agents with nearly *any* primary goal.

6. AI Agents' Interactions with Humans

We can make some confident predictions about how strategically human-level AI agents will behave with respect to humans.

a. **While humanity can still meaningfully affect its plans, a strategic AI system will aim to appear convincingly aligned with human goals, or incapable of harming humans, whether it really is or not.** We already see a smaller but related problem with "sycophancy" in LLMs, where the system will tell a user what the user wants to hear, and "sandbagging" in AI evals, where AI systems that can tell they're being evaluated will act less capable than they really are.

b. **Strategically human-level AI systems can use humans as actuators.** Nascent, misaligned, strategically human-level AI systems will likely need humans as physical actuators for a few months or years, and so will initially ensure that humans aren't broadly inclined to coordinate in shutting them off or opposing them. During this time, humans would happily build whatever the AI systems need to replace us as actuators.

c. **Powerful AI agents won't want to coordinate with agents with much less strategic power, and won't share our moral and ethical systems.** By default, if there are lots of strategically superhuman AI systems, they will want to coordinate with each other rather than with us (though they will temporarily prefer to benefit from us rather than killing us immediately). As a weak analogy, humans don't trade with or enfranchise chimps, elephants, dolphins, or cephalopods: the rule of law applies only to humans.

7. Current civilizational response

How is humanity currently responding to the relevant features of AI?

a. **Current concrete proposals aimed at improving AI safety do not address the core difficulties of the alignment and control problems in strategically human-level AI agents**. All proposals we're aware of thus far seem more likely to produce AI systems that *appear* to be aligned with human interests, than to provide much confidence that this appearance is borne out in fact. No proposals that we're aware of seriously engage with the inner alignment problem - most don't even seriously address the much easier outer alignment problem.

8. Policy requirements

We don't know exactly what policy responses, if any, will be sufficient to avoid catastrophes. But we can say a few things about the *necessary* aspects of policy responses.

a. **Racing for more-capable AI systems is incompatible with prioritizing the safety of those systems.** The problem being that, by default, we won't be able to tell that our AI systems have crossed the catastrophe threshold until it is too late.

b. **Current security measures at AI labs are grossly inadequate to protect against espionage and theft of critical AI breakthroughs, let alone AI insider risks.** The U.S. government will likely need to be heavily involved here, to manage security and safety risks.