Preamble: My overall belief in each point isn't necessarily in the same direction as the arguments I gave, as these were partly generated by noticing what arguments were missing from the existing list.

- 1a. Human intelligence isn't near any physical limits
    - Against
        - Between human cognitive enhancement and especially use of AI tools, humans
- 1b.
    - Against
        - An agent that is good at strategy will not necessarily not want to escape your control, especially if we design it. There is no law that links skill at strategy and manipulation to a specific desire to impose one's will in a way that conflicts with yours. Strategy is often best *demonstrated* in adversarial contests, but this just biases our observations to associate strategic agents with adversarial actions.
            - It's not very hard to keep top Scrabble player Nigel Richards from being the World Scrabble Champion, because he is a COVID-cautious recluse whose life goal is to bike across India. He has skipped several world championships because of infection risk, and if that weren't enough to tip the scales you could probably offer to support his bike trip.
            - Current narrow AIs only beat humans at e.g. videogames when they are specifically trained to play them, and general AIs when they are specifically prompted or trained.
- 1c.
    - For
        - Assuming the misaligned AI is already the size of a large nation and wants to grow its economy, there is a long list of other resources the AI would take from us if it could, and which would endanger most or all humans:
            - Industrial capacity. Robotics will be key to any of the AI's goals as its primary means of interfacing with the real world. Factories, oil refineries, blast furnaces, and the like are highly capital-intensive, and it is in the AI's interest to seize them from humans. The carrying capacity of Earth without any industrial machinery will be a small fraction of the current population and many will starve.
            - Arable land. Land is only useful for growing food because it receives sunlight. The 20-40% energy efficiency of solar panels is orders of magnitude higher than food crops' <1%, so not only will the AI want to take all of our land, it will make better use of it.
            - Clean air. Unless they depend on humans, AIs will not mind filling the environment with any of the trillions of toxic chemicals that don't affect their robot and silicon bodies, e.g. lead.
- 2a/2b
    - Against
        - Even without a fully mechanistic theory of how every part works, humans tend to gain experience with complex systems that makes them much more reliable.

        - Airplanes and the aerospace industry are actually an example of this. We don't primarily make systems safe by modeling every possible outcome in their behavior, but rather through trial and error combined with careful root cause analyses. Air travel is inherently extremely dangerous, but large airliners almost never crash, with a fatality rate of one in several million.
        - In contrast, NASA commissioned reports to estimate the safety of the Space Shuttle by modeling combinations of potential failures, and went through an elaborate process to ensure their millions of lines of code was bug-free by documenting and understanding every change. Nevertheless, two shuttles were lost with all crew, a fatality rate of 2.8%.
        - This pattern appears with other technologies too. Industrial farming is possible because we have so much experience and high-level understanding of how animals work, despite animal biology being nowhere near as well understood as physics. Weather forecasting is a science even though no one has solved the Navier-Stokes equations.
        - LLMs are about to become some of the most widely used technologies, and as such they are on pace to become some of the most reliable and well-understood. In addition, since they fully live in software, failures can more easily be reproduced and a root cause analysis much more easily done than for airplanes. More complex and capable systems are inherently harder to understand, but our experience with earlier AIs will apply to later ones, and this can hugely outweigh the complexity factor, much as a modern airliner is hundreds of times safer than a WWI fighter plane despite having >100x as many parts.
- 3a
    - Against
        - The amount of compute available strongly contributes to algorithmic progress by determining the number and scale of large experiments researchers can run. Therefore, limiting compute will slow down both main inputs to increased capabilities.
        - Historically, most technologies that undergo many orders of magnitude improvement rely on several sources of improvement, each of which have diminishing returns.
            - E.g. modern weather forecasting requires compute, advanced algorithms, and a global network of sensors, and if any one of the three were stuck in 1950 or 1960, weather forecasts would be much worse.
            - Likewise, modern chips need both advanced lithography nodes and modern design tools
- 3b
    - For
        - AIs have saturated basically all benchmarks we are able to construct, including intelligence tests and databases of graduate-level domain-specific questions. The Center for AI Safety is currently constructing "Humanity's Last Exam" through crowdsourcing; experts in every field get a reward of up to $5000 per question successfully submitted, and many questions are already too easy and are

rejected from the benchmark because current AIs can solve them. AI agents have also been playing at expert human level at the strategy game Diplomacy for almost two years as of 2024, so there's no reason this general intelligence wouldn't transfer to strategic acumen.

- 3c
- 3d
    - Against
        - Just because a technology is hugely impactful does not mean people race for its strategic advantages. See this list of technologies: https://chatgpt.com/share/66fe5938-09ec-800a-b335-aa84383dfa5b
        - If a technology is shared openly by academics (like lithium batteries), has primarily civilian rather than military uses (like vaccination), and has large benefits uncapturable by whoever invents it (like the Internet), an arms race is unlikely.
            - Currently closed labs are at the cutting edge of AI, but except for their scale, their models are less than a year ahead of open-source models. Innovations are published in one of the fastest-moving and cosmopolitan scientific fields.
            - Despite all the hype about military AI, most of the AI market is civilian, and I expect it to stay this way. The number of military and civilian applications is both large for sufficiently advanced AI, but civilian applications incentivize people and companies to distribute the technology, reducing the impact of a race.
            - The amount that people use LLMs is disproportionate to the actual revenue captured by AI companies. Combined with the effect of open-source models, it is not clear that future AI firms will get a big strategic benefit if it's only months before the technology to automate basically every job is open-sourced.
- 3e
    - For
        - Under the current model release cycle, models make huge leaps in capabilities every generation, often surpassing human performance in one generation. GPT4 got 90th percentile on the bar exam while GPT-3.5 was only 10th percentile, the same was true of intelligence tests. If this trend continues, no recursive self-improvement will be required for AI to overtake humans in other domains within months.
- 4a
    - Against
        - Many papers showing AI application to science are overstated, e.g. the big DeepMind materials science paper was not actually much novel science
- 5b
    - Against
        - Most arguments for this claim rely on the assumption that the AI is maximizing a utility function over final states of the world, which we need to either directly

specify or train into the AI, after which we just let the AI run. These assumptions are not met by current or probably future AI systems for a variety of reasons.
        - Alex Turner argues against this framing here
        - Utility maximizers' theoretical appeal is that they don't pursue dominated strategies. But agents with incomplete preferences are coherent in the same sense, and shutdownable.
        - Rather than trying to instill goals and letting the AI run, we can observe agents over time and iteratively shape their goals through further training, thus catching flaws we didn't notice at first. The initial agents don't need to be superintelligent.
- 5c
    - Against
        - We don't need to fully understand human values and goals if we construct AIs that do not desire to control the universe, are broadly and myopically helpful/harmless/honest, and optionally satisfy other properties like following laws. Such an AI would let humans stay in control and continue to make moral progress. These myopic goals are easier to verify than deep philosophical concepts, because we can observe whether the AI follows them in benchmarks and real life.
- 5d
    - For
        - Current overoptimization literature suggests it is basically reasonable to model current AIs as selecting random plans from a base distribution (e.g. pretrained model) and conditioning on success / high reward at some task. If this continues to be true, power-seeking actions (that is, actions that have a lasting increase in the AI's action space) will be incentivized so strongly that even specific training against those actions is insufficient to prevent them, because plans that include power-seeking will be more likely by huge factors like 2^1000.
    - Against
        - Thornley's incomplete preference proposal mentioned earlier mitigates the incentive for self-preservation
- 6a
    - Against
        - It will be difficult for a strategic AI to hide its goals like this while remaining misaligned, for several reasons.
            - Current AIs are far smarter when reasoning in chain-of-thought than in a single pass. If there is explicit instrumental reasoning about appearing aligned to serve other goals, we can simply read that chain-of-thought to expose it. Eventually, AI will be able to do complex reasoning in a single forward pass, but the most intelligent systems will always chain-of-thought.
            - AIs acting aligned for instrumental reasons will need to convincingly pass a huge range of safety checks in simulation, including simulations that no one is watching, changes to their code that make them less smart, and

other AIs using far more thinking time to probe for signs of misalignment. This is like a spy trying to avoid being exposed against a team of almost-as-intelligent spies who can give him any psychoactive drug, invent any situation and analyze his responses for years afterwards, and even differentiate through his brain to make him more likely to say certain things.
        - If the AI is always pretending to be aligned and benign, then the misaligned goals are basically vestigial because the AI can act the same way while actually being benign. Even if they are 80% pretending and 20% scheming, there is nothing actually reinforcing the misaligned goals, so training will tend to remove misaligned goals. (Humans have goals that persist on their own due to biological drives and an internal reward system, so this only applies if reward continues to be external)
- 6b
    - Against
        - By the time we have AI systems that use humans as their servants or employees or whatever, most companies and governments will already have AI strategic advisors, narrowing the strategic gap.
- 7a
    - Several proposals, if successful, would make good outcomes much more likely.
        - AI control research as currently done by Redwood basically sidesteps both the inner and outer alignment problem by extracting useful work out of partially or fully misaligned AIs
        - Chain-of-thought faithfulness: Matt Macdermott has an argument here.
            - If we can trust the reasoning traces, they give us strong evidence about whether the model is aligned. If the model is not aligned and we are confident in faithfulness, we can also remove misaligned goals by targeting the step at which they appear (e.g where the AI says "I want to murder people so…") with training or various safety techniques, whether they arise from outer or inner misalignment.
        - Interpretability
            - With level 7 interpretability we can detect the formation of deceptive alignment, and probably intervene on it.
    - Also see Turner's argument against the outer/inner alignment framework here