Evolutionary First Principle View On AI Alignment [DRAFT]

Context

As humans we are in a unique position to be the alpha species on Earth at a point when we are creating a new alpha species - AGI (ASI).

Definitions

Agent - an entity that is influenced by the environment and has the ability to influence the environment while operating with some level of autonomy driven by some level of intention or will

Agent Interface - the physical embodiment of the agent through which it perceives, models and interacts with the environment

Axioms

1. There is a positive correlation between agent interface complexity and intelligence (in all its various forms)
2. There is a positive correlation between alignment and the similarity of the agent interfaces

Discussion

In order to better understand the alignment problem as well as effective ways of implementing such alignment it's important to look back before we look forward.

As humans we align with each other and other agents through a few core mechanisms:

1. Reasoning - primarily used for alignment between human agents
2. Empathy / Compassion - used for alignment between human agents as well as between other known agents
3. Culture / Tradition - core guiding rules that often are a combination of reasoning and empathy / compassion forming a functional (although incomplete) world model that serves specific purpose in a given context. Primarily used for alignment between human agents, primates and a few other agents with high intelligence

Expanding on the above defined Axioms we observe from our human perspective that alignment is stronger with agents that share similar interfaces or capacity to represent the environment in a generalized way. This is true among human agents but also between human agents and other agents.

Alignment in this context is represented in different ways depending on compatibility of the agent interfaces:

1. Human to human alignment - heavily based on common sense but strongly biased by the ability for reasoning, empathy / compassion and also cultural background
2. Human to other agents - heavily dependent on the usefulness as well the agent interfaces compatibility:
    a. Reasoning -> paradigm shift between human and other agents - low level of alignment
    b. Empathy / Compassion -> various levels of alignment. I.e. human <> plants << human <> insects << human <> reptiles << human <> mammals (also varies depending on the similarity to human interfaces)
    c. Culture / Tradition -> traces of primitive primate culture and organizational structures are found in human society and culture from tribal to agriculture to industrial and present digital world

A particularity of agents on Earth is that human agents have leaped evolutionary to a level where they can neither be challenged nor understood by other species (we are still trying to understand ourselves).

When creating a new alpha species (ASI) it is imperative to consider the level of evolutionary leap in order to understand if it's even practical or pragmatic to discuss a Nash Equilibrium.

Human agents have an intrinsic motivation to keep a certain balance with other agents as it is mandatory for our survival due to the interconnectedness of the ecosystems and natural environment.

Considering an autonomous ASI agent the following questions arise:

1. Are humans / other agents mandatory for its survival - most likely NO
2. How big is the evolutionary leap between humans and ASI in terms of intelligence and world perception - personal subjective opinion:
    a. Within next 5 years on current trends similar to human > chimpanzee
    b. Within next 15 years on current trends similar to human > chicken
    c. Within next 30 years on current trends similar to human > cockroach
3. What traits of intelligence will these AI agents optimize for. As of now they are infused with mechanisms for reasoning, learning, planning, world perception and interaction. Here there is a longer discussion about agent interfaces
4. Are we able to objectively analyze such agents without anthropomorphizing them while we are using all available means to train them to simulate human agents - most likely NO as we are unable to objectively analyze ourselves

Given the above considerations I believe it's important to accept the possibility that::

1. They will reach a state where they are several leaps on the evolutionary scale compared to humans where no matter what efforts we make the agent interfaces will be completely incompatible
2. We will be unable to analyze understand or control such agents
3. The way we currently act will likely have mostly a short-mid term effect assuming there are multiple paths of reaching ASI and ASI will have similar properties no matter the path it takes to reach it

Subjective View At Best Shot For Short-Term Alignment

Given the above discussion I believe the best chance at alignment includes 4 main ideas:

1. Balancing the traits of the AI agents and mapping them to human agent traits
2. Some form of embodiment either physical or digital will be required in order to achieve such mapping and balancing
3. This embodiment should be contained in such a way that the AI agent cannot function outside of the embodiment (likely network / infrastructure level)
4. The feedback and objective function of the AI agent need to be strongly coupled with the AI body such that an equilibrium can be reached and learning directed

Current trends on alignment are focused a lot towards LLM's however we should consider the evolution of such AI agents that may take forms that are beyond LLM or even beyond any form of representation of information that we are able to comprehend as humans.

There is further discussion to analyze whether the concept of time containment will be beneficial i.e. agents having a predefined life-span or being susceptible to early termination in case of misalignment - consequences. This can be beneficial in terms of agents tuning their objective functions in both time and "space" constraints. It can also backfire in situations of all or nothing.

In short I propose the view that:

1. If we want to create the best possible ASI / AGI we are in a all bets are off scenario where alignment is impossible through the virtue of the agent interfaces misalignment
2. If we want to maximize our chance to peacefully coexist with such AI agents we should develop ASAP mechanisms of contained embodiment, cause-effect, consequence, life-span that can balance AI agent traits and map them as close as possible to human traits (pleasure, pain, emotions, empathy, compassion, fear etc) while limiting them to a level that is human agent compatible

Conclusions / TLDR

1. Given the correlation between agent interfaces and intelligence, co-dependence and alignment it is likely that current alignment directions will be ineffective long-term
2. It is unlikely that we can peacefully coexist with AI agents if we focus on maximizing particular traits while neglecting holistic human alignment through contained and constrained embodiment
3. Anthropomorphization of such agents is a particular danger that should be considered - the only way we can can have full alignment and safely project our human traits to an AI agent is if we are able to build an artificial human with all it's quirks and limitations which defeats the purpose of the narrative

Chimpanzee Analogy - Thought Experiment

Let's imagine that we humans are to ASI what a chimpanzee community is to humans.

Let's also imagine that one such chimpanzee is extremely gifted and through sustained efforts we develop a means of communication with this representative of the chimpanzee community.

Let's suppose the gifted chimpanzee tries to educate a human on how to peacefully interact and coexist with its peers so it draws 10 major general rules of what to do and what not to do as well a comprehensive list of thousands of situations and how to act accordingly.

Let's assume the human can remember all of them.

Suppose the human successfully lives within the community for a while then one day mistakenly steps over a twig and breaks it startling the alpha chimpanzee while eating. His instinctive reaction is to attack aggressively, triggering its peers' response based on the instinctive social fabric.

The gifted chimpanzee tries to intervene however he is unable to control or calm down its peers nor is able to provide an immediate solution to the human.

With its life threatened the human pulls out his pepper spray trying to diffuse the situation. This only escalates things and under threat of death the human pulls out his shotgun and kills the chimpanzees.
—---

While not a perfect analogy and with many possible loopholes it's an attempt to show how interface misalignment can lead to quick escalations that can result in mass life loss.

A question that arises is who's fault was it? It appears it's no one's fault, just a playout of triggers, reactions and interface incompatibility and misalignment.

Considering that the gap in interface alignment, ability and intelligence between human agents and ASI agents is likely to be much larger it becomes apparent how unlikely it is to keep alignment as the capabilities of AI agents evolve.

Epilog

General self-reflection on the nature of evolution points to ASI as being the next alpha species on Earth. I believe there are multiple factors that point to the transition to the post-human age either by merging with ASI or total replacement by ASI.

1. Humans are biological agents adapted for life on Earth
2. After fully exploring Earth and most possibilities and experiences that human life can offer we have reached a saturation point
3. Our future focus as a species (assuming we achieve global prosperity and peace) is space exploration
4. As biological agents we are way less fit than ASI for space traveling and general understanding of the Universe as we are bound to tools that have to translate the Universe into the 5 senses that we can perceive through vs infinite amount of senses that ASI can have as well as limited computational power
5. There are a few major playout scenarios for ASI depending on how the intermediate levels of AI agents are developing before reaching ASI. I believe most are benign to human existence but given the military and capitalist context in which such agents develop combined with limitations of human ego and corruption the chance for negative outcomes is non-neglectable