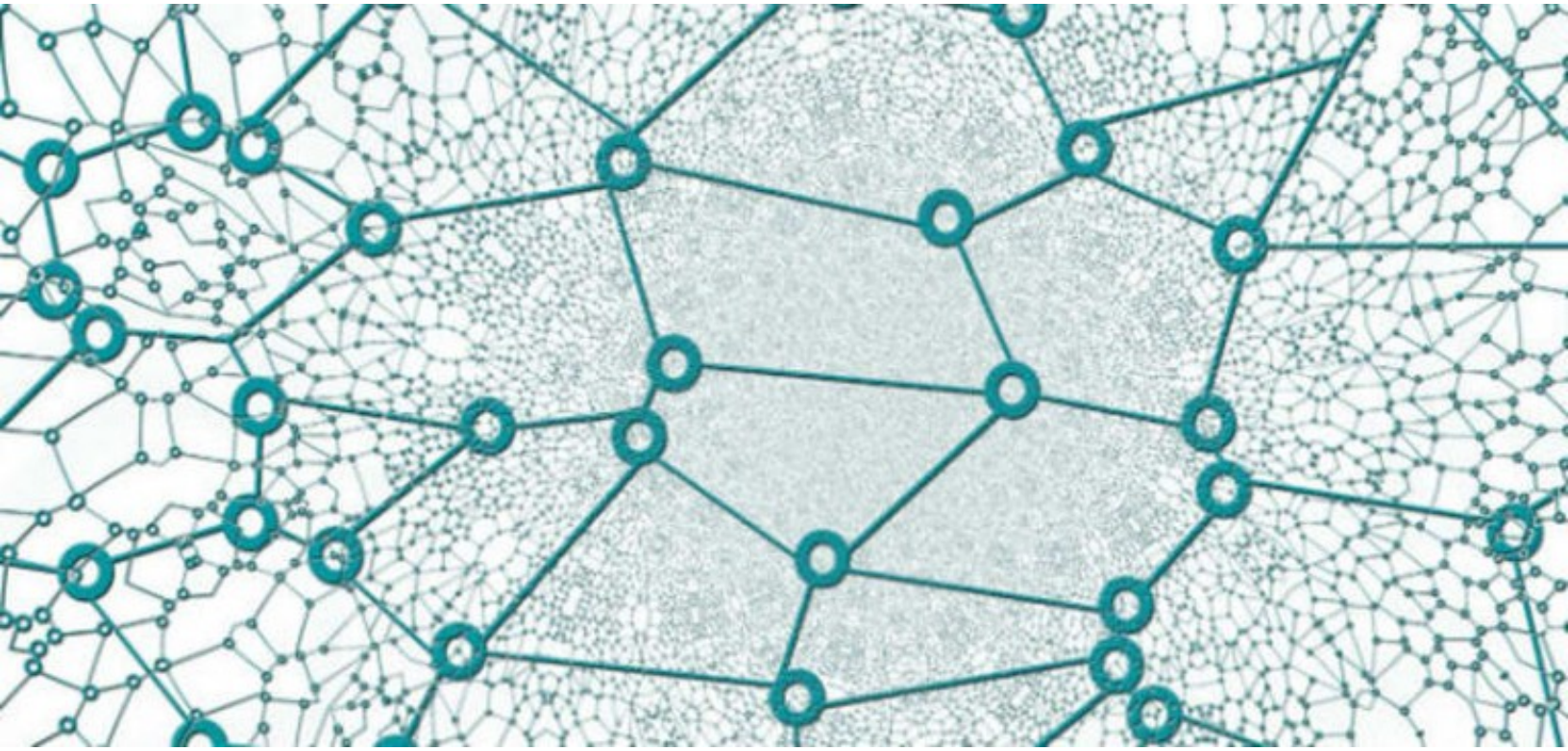

Some comments on the CAIS paradigm

1. The economy as an analogy for advanced AI systems

The economy as an analogy for advanced AI systems

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Eric Drexler's Comprehensive AI Services (CAIS), particularly as set out in his 2019 report *Reframing Superintelligence*, is a complex model with many different assumptions and implications. It's hard to grasp the whole thing at once, and existing summaries are brief and partial.[1]

One way of trying to understand CAIS is to seek generative intuitions for the whole model. These intuitions will be imprecise, but they can also make it easier to see why one might end up thinking that something like CAIS made sense.

In this post, we offer one such generative intuition for CAIS: using the economy rather than rational agents as an analogy for thinking about advanced AI systems.

Note that:

We are not making the historical claim that thinking about economies was in fact the main generator of Drexler's thinking on CAIS.[2]

There are other generative intuitions for CAIS, and other bodies of theory which the CAIS model is rooted in.[3]

The basic analogy

An economy is an abstraction for the sum total of 'the production, distribution and trade, as well as consumption of goods and services'. Prescriptively, we want the economy to serve human needs and preferences - and it does this at least to some extent.

Prescriptively, we also want advanced AI systems to serve human needs and preferences.

In worlds where we get advanced AI systems right, they would therefore be serving a similar function to the economy: serving human needs and preferences.

Whether we get AI right or not, it seems likely that advanced AI systems will become heavily integrated with the economy, such that it might become hard to distinguish them.

It therefore seems reasonable to imagine advanced AI systems in analogy with the economy, and to use what we know about economic dynamics to reason about dynamics which might shape those systems.

In the modern economy, specialised services are provided by a range of entities, mostly companies and governments. We don't see one giant global monopoly providing all services.

Thinking analogically about advanced AI systems, the CAIS model expects an array of specialised AI services working on decomposed tasks, rather than a single generally superintelligent agent (a global monopoly in the base metaphor).

This can be further unpacked. The reason that the human economy isn't structured as a global monopoly is that specialisation is efficient. It's often cheaper for an organisation to outsource a particular service, than to develop that capability in house: imagine family run businesses trying to manufacture their own smartphones from scratch, or big companies all hiring software engineers to develop their own internet search engines. So we end up with a range of different companies providing different services.

Note that specialisation isn't always the most efficient thing, because of economies of scope: cases where the unit cost of producing something decreases as the variety of products increases. Maybe you've already built a petrol station to sell petrol, and selling snacks too is cheap at the margin. Or you have a factory which makes women's shoes, and starting a men's line is pretty efficient. Here, there are joint costs which get shared across the different products, and so you get economies of scope.

But economies of scope don't seem to apply across the whole economy - otherwise we'd see the giant global monopoly. (Part of the reason here is that coordination costs increase with the size of an organisation, such that decentralised ways of sharing information like price signals are more efficient than centralised information transfer.) [4]

In the advanced AI analogy, Drexler argues that it will be more efficient for a given AI service to coordinate with other specialised AI services, than to learn to complete all tasks itself.

Some comments on related lines of thinking

- **Tool AI**: it's commonly argued that tool AIs want to be agent AIs. As we see it, the CAIS response to this argument is something like: maybe they do, but acquiring marginal agency isn't free. Drexler thinks that specialisation, task decomposition, and coordination with other systems will tend to be more efficient than becoming more agent-y. A possible response to this response is that beyond a certain capability threshold, it might suddenly become possible to reap economies of scope that were previously unattainable. We discuss this further in part 2 of our series.
- **RAAPs (Robust Agent-Agnostic Processes)**: the 'Production web' stories in What Multipolar Failure Looks Like envision a world where advanced AI systems look analogous to an economy. One way of looking at the relation between the Critch's production web and Drexler's CAIS analogy between AI systems and the economy is that they are thinking about similar take-off scenarios in different contexts. The production web stories were written in the context of exploring threat models, whereas Drexler's CAIS work is written in the context of exploring design space. Critch's Tech company singularity post is also drawing on a similar take-off scenario in the context of exploring policy options.
- **Interpretability and the Natural Abstraction Hypothesis**: following the economic analogy suggests that advanced AI will have some level of modularity. If this were the case, it might help us with interpretability. The intuition here is that rather than one black box, we'd have many black boxes, and the connections between the boxes might be transparent to us, allowing us to

understand more about how the system is working. But it's important to note that this depends on the modularity being legible and accessible to humans. If it is (perhaps because the natural abstraction hypothesis is true, or because of successful human efforts to design legible modularity), then modularity should make AI safer. If the modularity is not legible to humans, then CAIS may not be much easier to interpret than a monolithic AI agent.

Applications of the analogy

This economic analogy extends widely through CAIS. You could try to summarise the entire CAIS argument in terms of economics, but that would be hard. Here we do something easier, and just give some central examples where the analogy between the economy and advanced AI ecosystems applies.

Intelligence explosions

The view which Drexler is reacting to is something like:[5]

> A (single) AI system at a certain level of capability will converge upon certain instrumental goals, for a wide range of starting goals. One such instrumental goal is cognitive enhancement. Such an AI system would therefore recursively improve its own design, potentially leading to an intelligence explosion.

The CAIS model frames recursive improvement not as the recursive self-improvement of a single agent, but as continuous reinvestment in R&D:

> An ecosystem of AI services will include specialised AI services which perform automated AI R&D services. Continuous reinvestment in these R&D services leads to "asymptotically recursive improvement of AI technologies in distributed systems"[^6]: incremental research automation accelerates automation itself, potentially leading to an intelligence explosion.

The relationship between humans and AI system(s)

If you're centrally thinking about AI systems as agents, then the natural way to think about the relationship between humans and AI system(s) is in terms of principles (humans) and agents (AI system(s)).

In the CAIS model, the natural way to think about this relationship is instead in terms of consumers (other AI services, and ultimately humans) and producers (AI services).

It's not clear whether this distinction is important theoretically: it's possible that at some level of abstraction, both forms of economic relationships can be reduced to another. Conceptually, the distinction could be helpful for thinking about multi/multi alignment problems.[6]

Other applications of the analogy

- An application to AI safety approaches: existing economies use regulatory tactics to avoid externalities. In the CAIS model, security services (oversight-

based dynamics in the ecosystem) serve the same purpose. Misaligned companies are replaced with misaligned AI systems, and governments are replaced with a network of systems providing validation, monitoring, audit and other security services.
- An application to sources of AI risk: existing economies allow for activities aimed at the manipulation of consumer choice. Drexler talks about this in terms of 'unconstrained seductive and addictive services'; this maps approximately onto discussions of reward hacking in traditional AI safety circles.

There are probably lots of other applications too, though we think that intelligence explosions and the relationship between humans and AI system(s) are the most important.

The thinking behind this post has been spread out over several years, many people have contributed to it, and we are definitely missing some of those people in the following list. That said, we would like to thank: various FHI research scholars for comments on an early draft; VojtaKovarik , Cara Selvarajah and Nora_Ammann for comments on a subsequent draft; Fin Moorhouse and adamShimi for comments on the present draft; and Eric Drexler , Chris van Merwijk and Jan_Kulveit for numerous conversations.

Notes

The main summaries we are aware of are Rohin Shah's Reframing Superintelligence: Comprehensive AI Services as General Intelligence and Richard Ngo's Comments on CAIS. There is also a short review of _Reframing Superintelligence _by Scott Alexander. ↩

Though we do believe that thinking about economies was part of what generated CAIS. See e.g. Mark Miller and Eric Drexler, Markets and Computation: Agoric Open Systems (1988). ↩

For example, Jan Kulveit's take:

- "CAIS is a perspective which is rooted in engineering, physics and continuity"continuum"
- Agent foundations feel, at least for me, more like coming from science, mathematics, and a "discrete/symbolic" perspective".

An important question is whether this situation would hold for arbitrary levels of intelligence. We touch on this further in part 2 of the series. ↩

In this post we're not going to take a position either way on whether this is an accurate representation of others' actual views, or on who held or holds those views. ↩

Andrew Critch and David Krueger, *AI Research Considerations for Human Existential Safety (ARCHES)* (2020). ↩