Cyborg Periods: There will be multiple AI transitions

Jan Kulveit, Rose Hadshar

It can be useful to zoom out and talk about very compressed concepts like *'AI progress'* or *'AI transition'* or *'AGI timelines'*. But from the perspective of most AI strategy questions, it's useful to be more specific.

Looking at all of human history, it might make sense to think of ourselves as at the cusp of an AI transition, when AI systems overtake humans as the most powerful actors. But for practical and forward-looking purposes, it seems quite likely there will actually be multiple different AI transitions:

1. There will be AI transitions at different times in different domains
2. In each of these domains, transitions may move through multiple stages:

Stage [>> = more powerful than]	**Description**	**Present day examples**
Human period: Humans >> AIs	Humans clearly outperform AIs. At some point, AIs start to be a bit helpful.	Alignment research, high-level organisational decisions…
Cyborg period: Human+AI teams >> humans Human+AI teams >> AIs	Humans and AIs are at least comparably powerful, but have different strengths and weaknesses. This means that human+AI teams outperform either unaided humans, or pure AIs.	Visual art, programming, trading…

AI period: AIs >> humans (AIs ~ human+AI teams)	AIs overtake humans. Humans become obsolete and their contribution is negligible to negative.	Chess, go, shogi…

Some domains might never enter an AI period. It's also possible that in some domains the cyborg period will be very brief, or that there will be a jump straight to the AI period. But:

- We've seen cyborg periods before
  - Global supply chains have been in a cyborg period for decades
  - Chess and go both went through cyborg periods before AIs became dominant
  - Arguably visual art, coding and trading are currently in cyborg periods
- Even if cyborg periods are brief, they may be pivotal
  - More on this below

This means that for each domain, there are potentially two transitions: one from the human period into the cyborg period, and one from the cyborg period into the AI period.

Transitions in some domains will be particularly important

The cyborg period in any domain will correspond to:

- An increase in capabilities (definitionally, as during that period human+AI teams will be more powerful than humans were in the human period)
- An increase in the % of that domain which is automated, and therefore probably an increase in the rate of progress

Some domains where increased capabilities/automation/speed seem particularly strategically important are:

- Research, especially
  - AI research

    - AI alignment research
- Human coordination
- Persuasion
- Cultural evolution
    - AI systems already affect cultural evolution by speeding it up and influencing which memes spread. However, AI doesn’t yet play a significant role in creating new memes (although we are at the very start of this happening). This is similar to the way that humans harnessed the power of natural evolution to create higher yield crops without being able to directly engineer at the genetic level
    - Meme generation may also become increasingly automated, until most cultural change happens on silica rather than in brains, leading to different selection pressures
- Strategic goal seeking
    - Currently, broad roles involving long-term planning and open domains like "leading a company" are in the human period
    - If this changes, it would give cyborgs additional capabilities on top of the ones listed above

Some other domains which seem less centrally important but could end up mattering a lot are:

- Cybersecurity
- Military strategy
- Nuclear command and control
- Some kinds of physical engineering/manufacture/nanotech/design
    - Chip design
- Coding

There are probably other strategically important domains we haven’t listed.
A common feature of the domains listed is that increased capabilities in those domains could lead to large increases in **power,** for the systems with those capabilities. It will sometimes be helpful to consider power in aggregate, so that we can make direct comparisons about the amount of power which is automated in a given domain.

Clearly, capabilities in these domains interact. In our view, people coming from different backgrounds often perceive large increases in power in *their* domain of expertise as the decisive transition. For example, it is easy for someone coming from a research background to see how *automated research abilities* could impact other domains. But the reverse is also true: automated powers of persuasion, or automated cultural evolution, would have a strong impact on research, by making some directions of thinking unpopular, and influencing the allocation of attention and minds.

Note that it isn't clear that the level of abstraction we've picked here is the right one. It's possible that even more granularity would be more helpful, at least in some situations. For all of the domains we list, you could think of sub-domains, or of particular capabilities which might advance faster or slower than others.

The order of AI transitions in different domains will matter

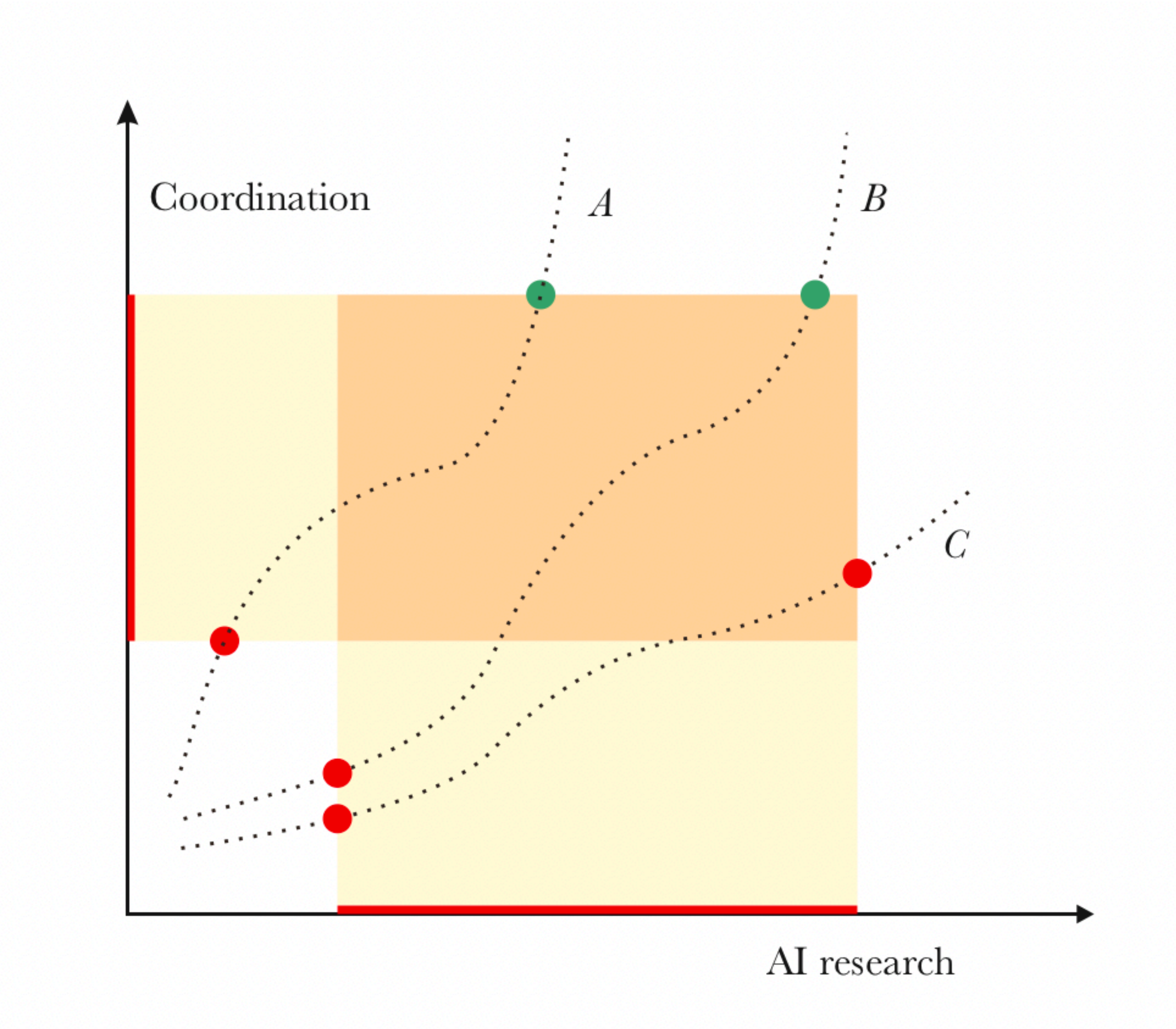
The timing of transitions in different domains isn't independent. But the world will look very different depending on which transitions happen first. A few vignettes:

- In a world where cultural evolution and AI research transition first, we may see the window of what's culturally possible opening fast and in unexpected directions:
    - Increasing the power of ideologies might cause leading AI research labs to become heavily regulated or nationalised
    - Concerns about AI sentience might become a large driving force behind AI research
    - In contrast, an ideology might emerge which promotes ceding power to AIs as virtuous and good
    - And many other possibilities (predicting future successful ideologies is obviously very hard)
- In a world where human coordination and manufacturing progress faster than other domains, humans might be able to leverage narrower AIs to bargain about the limits of power for AI systems deployed in socio-economic or political contexts, or about other aspects of AI development. Possibly, a "dominant coalition" could become powerful enough to enforce existential safety (c.f. Paretotopia).

Importantly, the fact that there are different possible orderings suggests that there are multiple possible winning strategies from the perspective of decreasing existential risk. For example:

- Moving faster on automating coordination than automating power is one possible route to minimising existential risk

- Moving faster on AI alignment research than AI research is another



Caption: in trajectories A and B, coordination is automated more quickly than AI research. In trajectory C, AI research is automated more quickly.

What does all of this imply? Tentatively:

- Actions that have the potential to differentially speed up automation in some areas over others could be very valuable.
- It seems unlikely that we will be able to accurately predict the trajectory we take in advance, with our current levels of understanding of the dynamics.
    - Insofar as we will have to rely on our ability to course correct rather than our ability to chart out the ideal trajectory ahead of time, becoming very good at course correcting seems desirable.

‘Cyborg periods’ could be pivotal

Even if cyborg periods are brief, they may be pivotal:

- Humans (via human+AI teams) will be more powerful actors than during human periods, and have more influence over future trajectories
    - This could be good, if the increases in power are directed towards risk-reducing things like coordination and alignment
    - It could also be bad, if the increases in power further exacerbate power inequalities between humans, aren't exercised with wisdom, are directed towards risk-increasing activities…
- It seems likely that the most important work for minimising risk will also happen during cyborg periods, because of increased power, and greater insight into what very advanced AI systems will look like
- Key deployment decisions will also probably happen during cyborg periods
- Once we enter AI periods where AIs are clearly more powerful than humans, it may be too late to change trajectories
    - This seems true at a general level
    - Whether it's true for particular domains probably depends on the ordering of AI transitions

This leads to a picture where there are overlapping but different cyborg periods in different domains. These periods will probably be:

- Weird: things that were impossible will be possible, rates of progress and change may be diverging significantly in different domains, the rules of the game will be changing
    - For the world as a whole to start feeling really weird, it's probably sufficient to enter the cyborg period in any of a small number of strategically important domains (research, coordination, persuasion, cultural evolution, probably a few other domains)
- High leverage: for the reasons above
- Fast-paced: it seems possible (though not inevitable) that cyborg periods will be short, and consequently feel like crises

Interventions

Leveraging the power of human+AI teams during cyborg periods seems like it might be critical for navigating transitions to very advanced AI.

This is likely to be non-trivial. For example, to really make use of the different kinds of cognition in a system involving a single AI system and a single human requires:

- Sufficient/appropriate understanding of the AI system's strengths and weaknesses
- Novel modes of factoring cognition, as well as means to implement a given factorisation, including e.g.
    - Specialised workflows
    - Good user interfaces
- Modifications of the AI system for this purpose

Doing this in a more complex set-up might involve a lot of substantive work. But we can probably prepare for this in advance, by practising working in human+AI teams in the sub-domains where automation is more advanced.

This applies more broadly than just to AI alignment research, and it would be great to have people in other strategically important domains practising this too.

The ideas in this text are mostly from Jan, and private discussions between Jan and a few other people. Rose did most of the writing. Clem and Nora gave substantive comments. s The post was written as part of the work done at ACS research group.