The SDGs of Strong Artificial Intelligence | Emerj

Daniel Faggella

While it is difficult for people to agree on a vision of utopia, it is relatively easy to agree on what a "better world" might look like.

The United Nations "Sustainable Development Goals," for example, are an important set of agreed-upon global priorities in the near-term:

Source: United Nations

These objectives (alleviation of poverty, food for all, etc.) are important to keep society from crumbling and to keep large swaths of humanity in misery, and they serve as common reference points for combined governmental or nonprofit initiatives.

However, they don't help inform humanity as to which future scenarios we want to move closer or farther to as the human condition is radically altered by technology.

As artificial intelligence and neurotechnologies become more and more a part of our lives in the coming two decades, humanity will need a shared set of goals about what kinds of intelligence we develop and unleash in the world, and I suspect that failure to do so will lead to massive conflict.

Before diving into the need for more long-term Sustainable Development Goals, I presuppose

the following hypotheses:

- In the next 20-40 years, much of humanity will become cognitively enhanced
- Humanity is reasonably likely to develop AI superintelligence within the 21st century
- Humanity, like all species before it, is on it's way to (a) extinction, or (b) transforming into something beyond itself (no species can remain unchanged forever)
- Humanity is amoral, and is incentivized to conflict whenever conflict serves its aims

Given these hypotheses, I've argued that there are only two major questions that humanity must ultimately be concerned with:

- What is a beneficial transition beyond humanity?
- How do we get there from here without destroying ourselves?

In the rest of this article, I'll argue that current united human efforts at prioritization are important, but incomplete in preventing conflict and maximizing the likelihood of a beneficial long-term (40+ year) outcome for humanity. I'll also pose some global priorities that might help to encourage this long-term outcome.

The SDGs of AI and Neurotech

The problem with setting goals with a longer time horizon than current SDGs lies in the fact that (a) the long-term future is incredibly hard to predict, and (b) unlike the near-term, there is no clear objective what a "better" long-term condition for humanity is.

It's also the cast that, for the most part, the SDGs address was of receiving suffering. Alleviation gender bias, alleviating education disparities, alleviating food shortages – the amelioration of suffering is hard to disagree on. Setting a transformative future vision that involves drastic change is much more challenging.

This challenge is necessary because it will provide meaning to why the SDGs themselves are relevant in the long-term. We might ask:In a hundred years, is the highest imaginable goal to have a planet of slightly happier human beings (and animals)?

- In a thousand years, what is the best possible world we could image?
- In other words, what is the long-term "point" of goals like the SDGs, what are they headed towards?

Many experts (link) agree that the next hundred years may lead to vastly different future scenarios:

- A future where human beings spend most of their time in immersive VR environments, hooked up to brain-computer interfaces that allow them to be vastly more productive and blissful at the same time (e.g. Lotus Eaters vs World Eaters)
- A future where superintelligent AI automates most economic activity and humans spend most of their time working on personally meaningful projects (charities, art, investing in relationships, etc)
- A future where artificial intelligence becomes conscious and superintelligent – and expands beyond Earth to populate the galaxy, turning entire planets into computing substrate to

expand its intelligence and understanding. If such a conscious superintelligent AI is unimaginably blissful (experiencing positive qualia to an almost unlimited degree). This may be among the best utilitarian outcomes (see "utilitronium shockwave")
- A future where human brain augmentation becomes globally banned, and humanity decides to limit the power of AI to ensure that humanity (as it is today in the early 20th century) remains the most advanced intelligence on earth for as long as possible.
- A future where some nations ban brain augmentation, and others promote it, and the human race splits off into a race of super-capable, super-intelligent humans, and "normal" humans, putting tremendous power in the hands of the enhanced group
- Etc…

Given the massive consequences of these objectives, it is likely that global conflict will arise as nations, political parties, and religious groups grapple with which futures it wants to orient itself towards.

All of these positive visions above have massive dystopian potential, and it's clear that the future poses many risks for human destruction or even extinction.

We are literally talking about determining the trajectory of intelligence itself, the intelligence that will potentially populate the universe, a leap in intelligence as grand as the leap beyond humans as humans were a leap in intelligence beyond bacteria.

For this reason, I posit that we need a set of long-term shared global goals:

Certainly we need efforts like the SDGs to improve the lives of this generation and the next, and to allow humanity to continue to thrive and evolve, but at some point we will need to come to an understanding of where we are ultimately headed, but we need something more.

Note: I'm using the United Nations' SDGs because they are a common touchstone of global priorities, not necessarily because the objectives I'm advocating need to be added to the existing SDGs or because the United Nations necessarily must be the vehicle through which the nations of the world come together to discern the futures they want to pursue.

Long-Term Goal 1: North Star Intelligence Trajectory Goals

I hold that the global community of humanity should work to develop near-term and long-term scenarios that we consider “probably beneficial,” and scenarios that we consider “probably detrimental.” There is no way to see the future, but there may be a way to agree more-or-less on scenarios that most (or many) humans think would be morally preferable or un-preferable.

There are no concrete endpoints that humanity can aim at. Change is constant, and there is no ultimate or singular destination. There are, however, “constellations” of possible future scenarios that humanity might choose to steer towards or away from.

Exploring Post-Human Intelligence With and Without International Coordination

We can imagine a world where each nation explores and/or invests in these “constellations” of futures in a disjointed and individual way. Some countries banning brain enhancement, some countries focusing on “upgrading” the minds of as many of their population as possible – some countries exploring AI superintelligence, other countries focusing on gene editing – some

countries focused on moving quickly to vastly post-human scenarios.

Such a disjointed international arena would inevitably lead to conflict.

On the other hand, we can imagine that there might be benefits to some degree of alignment between nations:

- No two nations would have to run the same experiments
- All nations and people would become more informed about potential future risks and opportunities
- There might be less conflict between nations who are pushing for or against certain future scenarios (gene-edited designer babies, for example)
- There could be a unified (hopefully rather democratic, or at least peaceful) process for course-correcting, and determining new futurescapes to target and explore

It's by no means certain that a unified (rather than disjointed) approach to moving towards a post-human future would be more likely to bring about better outcomes, but it seems likely to be that it would.

Long-Term Goal 2: Global Transparency and Steering of the Intelligence Trajectory

It seems somewhat inevitable that neurotechnologies and strong AI would require global governance.

If a unified council of nations determines the kinds of human brain augmentations that humanity should explore, and those which (at least for now, for the sake of safety) we should not explore, then somewhere and somehow, there would need to be a way to monitor the research and technology initiatives of companies and countries.

There are precedents for this kind of governance within countries (for example, in the USA there are many sectors which are regulated, and there are kinds of products that are illegal, or kinds of chemicals or components that require licenses to acquire), and there are some limited

precedents for this kind of governance between nations (nuclear proliferation has a kind of imperfect global monitoring process; there are chemical weapons agreements between many countries that allow for international transparency around the movement of certain dangerous materials). But what will be needed in the future will be an entirely new level of governance.

Transparency will be hard to arrive at because AI development is very hard to monitor. It may be unreasonable to trace and understand the code on all computers across the entire internet and still more unreasonable to get ahold of the content of computers not connected to a central network.

Proxies for different kinds of AI and neurotech research will have to be determined and monitored in order to ensure that strong AI or “unsafe for now” (as deemed by a unified international body) cognitive enhancements aren’t being built and released by rogue or revolting groups.

Steering will be required to align the world’s research initiatives to align with international priorities for post-human transition (and probably for other issues as well, like curing diseases and reducing pollution).

This would reduce duplicated research, improve global awareness of AI and neurotechnology initiatives, and ensure that the most potentially dangerous new research can be guided in what the international community believes to be the most aggregately beneficial near and long-term ends.

Without this kind of alignment (in both transparency and steering), it is reasonable to suspect that conflict would result.

A World of Post-Human Intelligence With and Without International Governance

An arms race would be the default state of affairs. Dominance by the most powerful nation or group would be the default state of affairs.

Global transparency and steering, I believe, would be among the only ways to reduce the

inevitability of war, and a way to escape an arms race dynamic.

Absolutely all of this is unreasonably optimistic to hope for, but the formation of the United Nations was deemed impossible by many, and all in all, it seems more than worthy to make the attempt to arrive at this kind of governance, even if it's remarkably hard.

Considerations for AI Power

The core hypothesis here is this:

> The selfishness of humanity and the competitiveness of nations will lead artificial intelligence into a devastating arms race unless humanity becomes united in an effort to determine what should be beyond humanity, and how we get there.

Assuming international coordination around the post-human transition becomes a viable path forward, companies and countries will jockey for power be advocating for governance models that behoove their own interests.

The interests of tech-powerful countries and companies (those with the greatest ability to develop powerful AI and neurotechnology advancements) will be very different from the less powerful countries and companies.

Here are some of the positions we might expect from more and less powerful countries and companies.

Power Dynamics

For the sake of this article we'll describe the "Tech Powerful" to be the companies and nations most likely to advance and profit from strong AI and neurotechnologies, and the "Less Powerful" as the companies and nations that stand little chance to advance or profit from strong AI and neurotechnologies.

International Governance

- Less Powerful: More likely to advocate for international governance.
- Tech Powerful: Will advocate against international governance, as they will benefit from a lack of such oversight.
- Tech Powerful: Powerful companies who see the tides turning towards inevitable governance may jump proactively into governance in order to gain a political and PR advantage over their competitors.

Membership and Participation

- Tech Powerful: will want some kind of permanent membership lock-in, like we see in the UN.
- Less Powerful: will advocate against having powerful permanent members dominate the organization, like the UN.
- Both: Will band together with countries that have similar aims in order to further their interests.

Speed and Method of Technology Development

- Tech Powerful: Might be more likely to advocate faster development as it will further their relative economic and military prominence.
- Tech Powerful: Will likely advocate to keep cutting edge initiatives in their own nations on the grounds of being most effective.
- Less Powerful: Might be more likely to advocate slower development so they aren't left behind.
- Less Powerful: Advocate for the distribution of some research initiatives to their home countries (instead of allowing the Tech Powerful countries to run them all) on the ground of fairness.

There is no way around conflict and competing incentives, international coordination or no international coordination. The hope, however, is that competing incentives can be sorted out through diplomacy and argument, rather than arms race and war. The hope is that foresight will allow us to structure international incentives towards collaboration (and away from conflict) before it's too late.