Growing Children For Bostrom's Disneyland

Posted on July 13, 2014 by Scott Alexander

Epistemic status: *Started off with something to say, gradually digressed, fell into total crackpottery. Everything after the halfway mark should have been written as a science fiction story instead, but I'm too lazy to change it.*

*
**

I'm working my way through Nick Bostrom's Superintelligence: Paths, Dangers, Strategies. Review possibly to follow. But today I wanted to write about something that jumped out at me. Page 173. Bostrom is talking about a "multipolar" future similar to Robin Hanson's "em" scenario. The future is inhabited by billions to trillion of vaguely-human-sized agents, probably digital, who are stuck in brutal Malthusian competition with one another.

Hanson tends to view this future as not necessarily so bad. I tend to think Hanson is crazy. I have told him this, and we have argued about it. In particular, I'm pretty sure that brutal Malthusian competition combined with ability to self-edit and other-edit minds necessarily results in paring away everything not directly maximally eco-

nomically productive. And a lot of things we like – love, family, art, hobbies – are not directly maximally economic productive. Bostrom hedges a lot – appropriate for his line of work – but I get the feeling that he not only agrees with me, but one-ups me by worrying that consciousness itself may not be directly maximally economically productive. He writes:

> We could thus imagine, as an extreme case, a technologically highly advanced society, containing many complex structures, some of them far more intricate and intelligent than anything that exists on the planet today – a society which nevertheless lacks any type of being that is conscious or whose welfare has moral significance. In a sense, this would be an uninhabited society. It would be a society of economic miracles and technological awesomeness, with nobody there to benefit. A Disneyland with no children.

I think a large number of possible futures converge here (though certainly not all of them, I myself find singleton scenarios more likely) so it's worth asking how doomed we are when we come to this point. Likely we are pretty doomed, but I want to bring up a very faint glimmer of hope in an unexpected place.

It's important to really get our heads around what it means to be in a maximally productive superintelligent Malthusian economy, so I'm going to make some assertions. Instead of lengthy defenses of each, if you disagree with any in particular you can challenge me about it in the comments.

- Every agent is in direct competition with many other entities for limited resources, and ultimately for survival.

- This competition can occur on extremely short (maybe sub-microsecond) time scales.

- A lot of the productive work (and competition) is being done by nanomachines, or if nanomachines are impossible, the nearest possible equivalent.

- Any agent with a disadvantage in any area (let's say intelligence) not balanced by another advantage has already lost and will be outcompeted.

- Any agent that doesn't always take the path that maximizes its utility (defined in objective economic terms) will be outcompeted by another that does.

- Utility calculations will likely be made not according to the vague fuzzy feelings that humans use, but very explicitly, such that agents will know what path maximizes their utility at any given time and their only choice will be to do that or to expect to be outcompeted.

- Agents can only survive a less than maximally utility-maximizing path if they have some starting advantage that gives them a buffer. But gradually these pre-existing advantages will be used up, or copied by the agent's descendants, or copied by other agents that steal them. Things will regress to the pre-existing Malthusianism.

Everyone will behave perfectly optimally, which of course is terrible. It would mean either the total rejection of even the illusion of free will, or free will turning into a simple formality ("You can pick any of these choices you want, but unless you pick Choice C you die instantly.")

The actions of agents become dictated by the laws of economics. Goodness only knows what sort of supergoals these entities might have – maximizing their share of some currency, perhaps a universal currency based on mass-energy? In the first million years, some agent occasionally choose to violate the laws of economics, and collect less of this currency than it possibly could have because of some principle, but these agents are quickly selected against and go extinct. After that, it's total and invariable. Eventually the thing bumps up against fundamental physical limits, there's no more technological progress to be had, and although there may be some cyclic changes teleological advancement stops.

For me the most graphic version of this scenario is one where all of the interacting agents are very small, very very fast, and with few exceptions operate entirely on reflex. It might look like some of the sci-fi horror ideas of "grey goo". When I imagine things like *that*, the distinction between economics and harder sciences like physics or chemistry starts to blur.

If somehow we captured a one meter sphere of this economic soup, brought it to Earth inside an invincible containment field, and tried to study it, we would probably come up with some very basic laws that it seemed to follow, based on the aggregation of all the

entities within it. It would be very silly to try to model the exact calculations of each entity within it – assuming we could even see them or realize they are entities at all. It would just be a really weird volume of space that seemed to follow different rules than our own.

Sci-fi author Karl Schroeder had a term for the post-singularity parts of some of his books – Artificial Nature. That strikes me as exactly right. A hyperproductive end-stage grey goo would take over a rapidly expanding area of space in which all that hypothetical outsiders might notice (non-hypothetical outsiders, of course, would be turned into goo) would be that things are following weird rules and behaving in novel ways.

There's no reason to think this area of space would be homogenous. Because the pre-goo space likely contained different sorts of terrain – void, asteroids, stars, inhabited worlds – different sorts of economic activity would be most productive in each niche, leading to slightly different varieties of goo. Different varieties of goo might cooperate or compete with each other, there might be population implosions or explosions as new resources are discovered or used up – and all of this wouldn't look like economic activity at all to the outside observer. It would look like a weird new kind of physics was in effect, or perhaps like a biological system with different "creatures" in different niches. Occasionally the goo might spin off macroscopic complex objects to fulfill some task those objects could fulfill better than goo, and after a while those objects would dissolve back into the substratum.

Here the goo would fulfill a role a lot like micro-organisms did on Pre-Cambrian Earth – which was also intense Malthusian competition at microscopic levels on short time-scales. Unsurprisingly, the actions of micro-organisms can look physical or chemical to us – put a plate of agar outside and it mysteriously develops white spots. Put a piece of bread outside and it mysteriously develops greenish white spots. Apply the greenish-white spots from the bread to the white spots on the agar, and some of them mysteriously die. Try it too many times and it stops working. It's totally possible to view this on a "guess those are laws of physics" level as well as a "we can dig down and see the terrifying war-of-all-against-all that emergently results in these large-level phenomena" level.

In this sort of scenario, the only place for consciousness and non-Malthusianism to go would be *higher level structures*.

One of these might be the economy as a whole. Just as ant colonies seem a lot more organism-like than individual ants, so the cosmic economy (or the economies around single stars, if light-speed limits hold) might seem more organism-like than any of its components. It might be able to sense threats, take actions, or debate very-large-scale policies. If we agree that end-stage-goo is more like biology than like normal-world economics, whatever sort of central planning it comes up with might look more like a brain than like a government. If the components were allowed to plan and control the central planner in detail it would probably be maximally utility maximizing, ie stripped of consciousness and deter-

ministic, but if it arose from a series of least-bad game theoretic bargains it might have some wiggle room.

But I think emergent patterns in the goo itself might be much more interesting.

In the same way our own economy mysteriously pumps out business cycles, end-stage-goo might have cycles of efflorescence and sudden decay. Or the patterns might be weirder. Whorls and eddies in economic activity arising spontaneously out of the interaction of thousands of different complicated behaviors. One day you might suddenly see an extraordinarily complicated mandala or snowflake pattern, like the kind you can get certain variants of Conway's Game Of Life to make, arise and dissipate.

Source: Latent in the structure of mathematics

Or you might see a replicator. Another thing you can convince Conway's Game of Life to make.

If the deterministic, law-abiding, microscopically small, instantaneously fast rules of end-stage-goo can be thought of as *pretty much* just a new kind of physics, maybe this kind of physics will allow replicating structures in the same way that normal physics does.

None of the particular economic agents would feel like they were contributing to a replicating pattern, any more than I feel like I'm contributing to a power law of blogs every time I update here. And it wouldn't be a disruption in the imperative to only perform the most economically productive action – it would be a pattern that supervenes on everyone's economically productive behavior.

But it would be creating replicators. Which would eventually retread important advances like sex and mutation and survival of the fittest and multicellularity and eventually, maybe, sapience.

We would get a whole new meaning of *homo economicus* – but also pan economicus, and mus economicus, and even caenorhabditis economicus.

I wonder what life would be like for those entities. Probably a lot like our own lives. They might be able to manipulate the goo the same way we manipulate normal matter. They might have science to study the goo. They might eventually figure out its true nature, or they might go their entire lifespan as a species without figuring out anything beyond that it has properties it likes to follow. Maybe they would think those properties are the hard-coded law of the universe.

(Here I should pause to point out that none of this requires literal goo. Maybe there is an economy of huge floating asteroid-based factories and cargo-freighters, with Matrioshka brains sitting on artificial planets directing them. Doesn't matter. The patterns in there are harder to map to normal ways of thinking about physics, but I don't see why they couldn't still produce whorls and eddies and replicators.)

Maybe one day these higher-level-patterns would achieve their own singularity, and maybe it would go equally wrong, and they would end up in a Malthusian trap too, and eventually all of their promise would dissipate into extremely economically productive nanomachines competing against one another.

Or they might get a different kind of singularity. Maybe they end up with a paperclip-maximizing singleton. I would think it much less likely that the same kind of complex patterns would arise in the process of paperclip maximization, but maybe they could.

Or maybe, after some number of levels of iteration, they get a positive singularity, a singleton clears up their messes, and they continue studying the universe as superintelligences. Maybe they figure out pretty fast exactly how many levels of entities are beneath them, how many times this has happened before.

I'm not sure if it would be physically possible for them to intervene on the levels below them. In theory, everything beneath them ought to already be literally end-stage. But it might also be locked in some kind of game-theoretic competition that made it less than

maximally productive. And so the higher-level entities might be able to design some kind of new matter that outcompetes it and is subject to their own will.

(unless the lower-level systems retained enough intelligence to figure out what was going on, and enough coordinatedness to stop it)

But why would they want to? To them, the lower levels are just physics; always have been, always will be. It would be like a human scientist trying to free electrons from the tyrannous drudgery of orbiting nuclei. Maybe they would sit back and enjoy their victory, sitting at the top of a pyramid of unknown dozens or hundreds of levels of reality.

(Also, *just once* I want to be able to do armchair futurology without wondering how many times something has already happened.)