filosofia unisinos | Unisinos Journal of Philosophy

ISSN 1984-8234

Filosofia Unisinos
Unisinos Journal of Philosophy
24(1): 1-10, 2023 | e24110

Unisinos – doi: 10.4013/fsu.2023.241.10

Interview

On Human Enhancement, Optimism, Risk, Existential Risk, and How to Manage and Regulate It: An Interview with Anders Sandberg.[1]

Sobre melhoramento humano, otimismo, risco, risco existencial e como gerenciá-los e regulá-los: uma entrevista com Anders Sandberg.

Murilo Mariano Vilaça
https://orcid.org/0000-0001-9720-5552
Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública, Rio de Janeiro, RJ. Brasil. Email: murilo.vilaca@fiocruz.br

Murilo Karasinski
https://orcid.org/0000-0002-6099-6968
Pontifícia Universidade Católica do Paraná, Educação Continuada da Escola de Educação e Humanidades, Curitiba, PR, Brasil. Email: k.murilo@pucpr.br

Introduction

In the text, we shared the interview with one of the leading scholars on some of the main subjects of the contemporary practical philosophical debate, Professor and researcher Anders Sandberg.

In 2003, he received his Ph.D in computational neuroscience from Stockholm University, Sweden, for work on modeling neural networks of human memory. Before that, in 1998, Sandberg had already participated in an important group of authors who crafted the first version of the well-known Transhumanist Declaration.

[1] As acknowledgments, we inform that Murilo Mariano Vilaça is supported by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). Grant Number: 201.377/2021. He is also a CNPq research fellow (APQ – PRÓ-HUMANIDADES).

Currently, Anders Sandberg is Senior Research Fellow at the Future of Humanity Institute (FHI), Senior Research Fellow at the Oxford Martin School, and Ethics and Value Fellow at the Reuben College, which are institutions of the University of Oxford. At FHI, he conducts research on management of low-probability high-impact risks, estimating the capabilities of future technologies, and very long-range futures.

His expertise extends to the following topics: The search for extraterrestrial intelligence (SETI), transhumanism, human enhancement, neuroethics, global catastrophic risks, future studies, and computational neuroscience.

He is the author of several highly relevant publications, some of them coauthored with other exponents of the debate. Given the magnitude of their contribution to the debates around their research topics, it would be a hard job to highlight the main ones. We therefore invite readers, according to their focus of interest, to search for the author's important contributions.

Besides scientific publications in neuroscience, ethics and future studies he has also participated in the public debate about human enhancement, existential risk and SETI internationally.

Despite his extensive curriculum and his importance to the scholarly and public debate, he is an extremely attentive and friendly person. We were able to see this at the CSER Conference 2022, an event organized by The Centre for the Study of Existential Risk (CSER) at the University of Cambridge. Prof. Sandberg and we attended the same session. After the presentation we were kindly questioned by Anders Sandberg, who focused on a central point of our argument. This highly appropriate intervention motivated us to invite him to give us an interview. To our surprise and honor, we were promptly answered, finding a space in his busy schedule. Less than a week later, the interview was done.

We hope that the initiated interlocution can be broadened and deepened soon with Anders Sandberg's coming to Brazil. This is the first in a series of interviews with leading international researchers conducted by the Philosophical Research Group on Transhumanism and Human Bioenhancement - GIFTH+ (Fiocruz/CNPq).

MK: Can I call you Professor Anders?

Anders Sandberg (AS): Yeah, you can do that. I don't mind.

MK/MV: All right. We are very happy and honored to have the opportunity to meet you and to be able to talk briefly about issues that we feel are extremely relevant to our days. We are organizing a special issue of the Unisinos Journal of Philosophy, and we would like to ask you some questions regarding existential risk and human enhancement, the role of technoscience and human extinction, preservation, and thinks like that. So before we start our specific questions about the theme of this special issue, I would like to know something about your career. You are a well-known thinker associated with transhumanism, being, for example, one of the signers of the Transhumanist Declaration. One of the criticisms I found about transhumanism is that it is too technologically optimistic, undermining the risk of human enhancement. So, also to reinforce the implausibility of this criticism, we would like to hear from you about how you came to the topic of risk, and to recover in general lines your trajectory in the studies about transhumanism, for example.

AS: I usually explain that my origin story is that I grew up in a very boring 1970s suburb outside Stockholm in Sweden, so I read all the science fiction I could find at the local library. At some point I decided to try to make it real, so I thought that I probably needed to learn the science. So I started reading all the science, and then I just kept on going. I have always been interested in the possibility of taking various fictional visions and making them real while assessing the likelihood of them becoming real. So I did my original academic studies in computer science and mathematics, got a PhD in computational neuroscience, where I studied neural networks, but at the same time I was working on setting up both a local transhumanist association in Sweden and participating in online debates back in the 1990s. After all, coming into the internet in 1991, that was a revelation. Before that, all discourse was local. Suddenly, I could participate in a global discourse with transhumanists. A bit later in the 1990s, I started setting up webpages explaining transhumanist concepts, which again became very important as vocal points. A lot of people found these webpages and said: "Oh, so that is what it's called. Now I know I'm a transhumanist!" So my trajectory was very much that I had this side project of transhumanist activism and organization, and my research about various emerging technologies, and gradually they became more and more interleaved, and eventually, as I came here to Oxford to the Future of Humanity Institute, which is part of the university of Oxford Philosophy Faculty, I actually completed the merge. So the interesting thing is that my academic trajectory has very much been a generalist one, so while I'm using computer science and mathematics as tools, I'm also very happy to study Psychology and Neuroscience and Philosophy, and then try to work with people who are experts in their fields to actually write good papers. So this is for example how some of my more successful papers have come about. I co-authored them with actual philosophers. After all, technically I am not a philosopher. I only have my high school Philosophy to my name, or I worked together with somebody who is an astronomer and know the other relevant part. So that is one way we can talk about this trajectory. Another interesting thing you mention is, of course, this issue of risk. And how optimistic is transhumanism? I think there are many different forms of optimism, too. So there is one form of optimism that I really have a hard time with, and that is essentially the deterministic assumption that everything is just going to become better and better. Sometimes I think this accurately describes Ray Kurzweil's take on, for example, in *The Singularity is Nearer*, where he more or less argues that there is an almost deterministic behavior of how technology and everything else is advancing, which means that initial decisions, individual decisions, don't matter. Randomness doesn't matter; it gets averaged out. Now this, I think, is wrong. I think it is a mistake to say that the world is like that because if we look around us, we see path dependencies all over the place. Individual decisions to do certain things or not do certain things have actually affected how the world functions. Indeed, there are desirable technologies that many pursue are very likely to come about, but not always because sometimes they are hard. Sometimes decisions go in other directions. We see a lot of it also in subtle ways like how software standards are formulated, but cannot be very long-term effect. So I would argue that there is a formal optimism that says that the future can be very good if we make it good – and/or lucky. So that means that the future is perhaps even likely to be good, but we need to work rather hard to ensure that it becomes good. Now, this can of course be applied on a global scale and a microscopic individual scale. One can think about one's own life: do you expect your career to go well or do you expect your career to go well if you work hard on your career? And most people will, of course, realize that they probably need to work somewhat well at their career to have it go well. Similarly, we might think about risks on an individual level. And I think most of us recognize that yes, if I eat healthy foods, exercise, and avoid doing very unhealthy things, I have a good chance of living a long and healthy life. But I can get run over by a truck, too. Even if I eat the healthiest food possible, that is no defense against accidents. So there is a continuous here. And this is where things get very interesting in relation to transhumanism and then later as we can discuss enhancement: transhumanism essentially argues that the future has a higher variance than most people expect the future to have. Most people expect the future to be like the present, but maybe flying cars, a bit more pollution, but that's about it. But transhumanists say that the future can become radically different. People normally tend to then focus on radically better. But if you're intellectually honest, you will realize that you could have radically worse, too. And this is why it is not a coincidence why so much work on existential risk was done by people who come from a transhumanism background. Nick Bostrom, who coined the modern discourse on existential risk, wrote his paper in 2003. I remember reading it on the way to meet with him at the Transhumanism Conference over at Yale University, and I was annoyed. I felt that it was going to be used as an excuse by luddites and governments who wanted to limit technological progress. What was he going on about? Over the years I've recognized that it actually has a very important point: if the world gets destroyed, then we are not going to get any of the positive aspects of a transhuman future. We're not even getting any future! So we have this interesting situation that while there might be what we can call deterministic or naïve optimists in the enhancement debate who think that everything is just going to get better, since they are ignoring the possibility of risk or even choice, they don't even need to make an argument, they should just expect everything to happen. A more realistic view is that we have choices, there are things that can go wrong, and if one takes the extreme transhumanist position, like things can be good or bad in extremely powerful ways, which means that we might actually need to be much more cautious, one of the paradoxical results is that the more you expect the future to be grandly good, the more you want to be really careful about getting there. If we were to expect the future to be just a bit depressive, in that case we might certainly want to get there, we want to avoid disaster, but we wouldn't have that urgency, that intensity of desire that we get if we thought the future were much more powerful. Similarly, there is an optimism about agency here. The transhumanist optimism about agency is that we can affect the future, which means that we also have reason to believe that we can reduce the risks. A fatalistic view of risks is that risks happen no matter what we do, and we individually or collectively have no influence. In that case, there is no point in doing existential risk research. Ok, so that was a micro lecture. Sorry about that. But I think it might be a starting point.

MK/MV: Yeah, excellent starting point. Actually, we have four questions, but the first question you already said a few things, so I will just copy and paste here. I guess it is better or you can just follow our idea. So our first question is: how do you evaluate the uses of the concept of risk in the human enhancement debate? In your opinion, has it been, as a rule, properly used? Or do you identify a relevant frequency of misuse of the concept, which would still hinder the creation of an adequate heuristic to deal with the challenges of human enhancement?

AS: Normally, when we talk about risk, people make it easy and say that it iss a combination of the probability of something bad happening and how much harm the bad is doing. And a lot of discussions happen about the probability, because we are uncertain about it, and one can argue both in a qualitative and quantitative way about it. And there are even interesting issues about how you combine them. But normally when dealing with risks like industrial accidents or car accidents or even existential threats to humanity, the kind of harm is usually taken to be unproblematic. But when we start thinking about harm in human enhancement, actually there are different, qualitatively different forms of harm. So the most basic one, which I think in many cases people are not really too concerned about, is medical harm. I put an implant in my body, I get an infection, I get harmed by that. This is not very different from any standard harm in medicine, and one can also think about it using standard medical issues. Is that form of harm a big problem? Well, it depends on what I'm attempting to do, but I don't think we have a problem in the debate about it. But usually in the ethics debate I've been involved in, there are of course various other forms of harm. So the most common idea is that it might be bad for autonomy. It might be that, if I get a brain chip implant, maybe my faults are not entirely my own, maybe I could be hacked, maybe there is a threat now to my autonomy, and this is a bad problem. Indeed, I wrote a paper about the issue of hackable neural implants, and I do think we have a real problem here in making them safe enough so people can trust them. But it's a harm that is very different from me suffering from an infection. Because even if no hacker ever hacks my brain implant, if I'm constantly second guessing my decisions because maybe somebody manipulated me, I'm suffering in a way. Even worse is that I might be actually manipulated, and there is a loss of agency here. But there are other forms of harm that people have been bringing up. Francis Fukuyama, in his *Our Posthuman Future*, was basically trying to use every anti-enhancement argument he could. But deep down, at the core, there was something impairing factor X he didn't quite say it, but it was basically what he was getting at. He was arguing that there is a human essence, a human nature, and this nature must not be changed, because then we lose something that defines us and is super important for us. That is, again, different from a lack of autonomy. Now we're getting to something that in a religious framework you would say that maybe this harms the soul directly. I think many secular thinkers would not accept that as an argument, and this is partially of course why Fukuyama chose not to really push the religious point and just hand way that there are certain things we don't want to harm. But now we get to very interesting matters. While most people would say that autonomy is important and we need to safeguard it, different cultures have different concepts of autonomy. A Chinese bioethicist would probably not regard individual autonomy as quite as important as an American bioethicist. And similarly, on being bad for the soul – that depends on what kind of soul you believe we have. So Fukuyma, he started from an Aristotelian essentialist view of the human nature, that we have an essential human nature and that we must remain true to. Many transhumanists, like me, always started on citing Pico della Mirandola's "Oration on the Dignity of Man" believing that God gave us the ability to change who we are. And that means that the changeable nature is actually what we need to safeguard. So something that makes us unchanging, oh yes that's a threat! But the things that allow us to change – that is a good thing. At this point, we end up with a risk concept that gets problematic and exciting in many ways because we have different qualitative forms of risk that might need to be traded against each other. And we might have disagreements about what constitutes a risk in this case. So I might want to get that brain implant acknowledging that there is some risk of getting hacked but it has abilities that might allow me to modify my own existence in various flexible ways. Fukuyama would say: "Anders, that is very bad, because now we are threatening something that is very valuable, and you don't even recognize that value". And I would of course argue back at him and say: "look, I kind of understand where you're coming from but I have a different value system". Now this is of course going to be a real tricky situation, because normally when dealing with risk, we're living in a risk-oriented society, and that usually means that people agree on risks. Most people assume that risks are [found], you can compare them, you can have greater risks, and smaller risks, and we can choose the smallest risk or calculate the risk reduction per dollar. When you have these multidimensional risks, it actually becomes a real ethical question, and there might not be any good simple answer everybody can agree on. This is maybe good news for a bioethicist who wants to remain in business, but bad news for making a simple decision, because this means that whatever, let's say, a doctor says it's ethical or not ethical to do in a surgery, we might disagree and say that actually we think that he is applying the wrong ethical framework in that case. I still think we can come up with a lot of agreements, and compromises are always possible, so it's not like everything is impossible. It's just that there is much more talking and discussing and disagreeing we have to do than just comparing some numbers.

MK/MV: Fascinating, professor Anders, and we could go on this discussion forever, I guess, but our time is limited, so I'll copy and paste the second question: regarding the concept of existential risk, which was formulated twenty years ago, how do you evaluate its development, the uses that have been made, and the criticism that it has received?

AS: Existential risk is interesting because it's intended to be qualitatively different from mere global catastrophic risks. So at some point a risk that is hurting a lot of people in the world goes from not just hurting the current victims but all of the future. It has an effect on the eventual outcome of the history of our species. Now, refining this has taken a lot of time. Nick's original idea got refined in a later paper, and many of my colleagues have been adding various thoughts about how to think well about it. For example, right now our standard definition is something like "something that really harms the potential of Earth originating intelligent life", because we might not want to be too anthropocentric. But "potential" here is a real philosophical problem. Do we know our potential? No, we don't. So how do we know whether something harms it? Again, getting back to the debate with Fukuyama: Fukuyama might say that the humanity that enhances itself might have lost a lot of its potential, even though his future super humans are going around doing a lot of super human stuff, but they lost their souls in some important sense. There have been other forms of criticism. The most annoying one is when people say that we are worried about far future stuff but there are many big important things here and now. Quite often you find this with people deeply invested in climate change. They either think that any other risks besides climate change take effort away from climate change, and climate change is the most important risk, so hence other risks can't be important, and hence you need to argue against it. I literally heard Steven Pinker make these arguments against taking artificial intelligence risks seriously, which is a horrendously stupid approach. Yes, there is a limited amount of public concern, but that doesn't mean that you can say that a risk can't be taken seriously just because we have a problem of focus. There is also a deeper philosophical question: how much do we care about the present generations, the nearby generations and the far future? How do you do that balancing in the right way? So there is an interesting debate of long termism, the view that most of the value resides in the far future, which I generally think I lean towards, and the view that if you take that too seriously, you would sacrifice the present for the future. Conversely, if we focus only on saving the present, we might sacrifice the future. There is a real justice question, also, about how we trade justice of future generations versus present generations. So I think another important criticism, which I find interesting even if I don't personally believe it is a strong one, is that besides existential risk, there might be suffering risks. It might be that animal suffering has moral weight just like human suffering. In that case, if we spread life across the universe, we might have endless amounts of very happy human civilizations, but we are also going to have entire planets full of jungles, full of wild animal suffering. We might have multiplied the total amount of suffering, which might count morally, and actually make the situation much worse than having just one planet with one suffering biosphere. So suffering risks could potentially be just as important as existential risks if you give moral priority to suffering. This is an important philosophical question we better figure out before we start spreading life in the universe. I think what is happening at present is that we are nuancing existential risk, we are recognizing that existential risks are among the key things, but there might be other things. Sometimes we joke that there might be an entire alphabet besides X risks (existential risks), and S risks (suffering risks). Maybe there are other ones that are just as important. We haven't discovered them yet. There might be ongoing moral catastrophes we need to overcome. However, they still seem to be really important, and are a very useful way of framing what research we can do. They are also very good for generating interesting hypotheses and challenges for decision making. How do you make decisions about the long term future? How do you handle the fact that some things seem to have a moral importance that are so many orders of magnitude greater than anything else? But it might be hard to balance the equation. Normally, when making choices, one side might weigh more heavily than another, not by a factor of a quadrillion. But in this case there might be these extremes, which make many people worry if it might not lead to fanaticism, a realization that anything is allowed in fighting existential risk. There is also an interesting question: when global risks become larger, and shade over existential risks, does anything change in how we should deal with them? For example, justice is a very important issue for global disasters. Generally, the poor and the weak get harmed the most, so we have a good reason to stop global disasters from a justice perspective, besides the fact that they hurt a lot of people. But as we become existential, we might end up with a situation where everybody dies, in which case justice is no longer part of it. And you might even say that if we manage to make an arc that saves some people, it might actually not matter if it's a very unjust system that saves them because justice requires survivors. You can't say that that was an unjust way of surviving and hence it would have been better if everybody died. At least most people would say so. But this of course goes back to the fact that a lot of this discussion has been framed, most of all from a consequentialist standpoint, rather than a deontological standpoint. And I think deontological existential risk studies are a very underdeveloped field that we might want to look into too.

MK/MV: Yeah, absolutely, so our third question is here
AS: Wow, that is a long one!

MK/MV: Yeah. In your opinion, at what stage are we in relation to the methodological challenges of risk analysis, especially the analysis of anthropogenic and intermediate existential risks (to use the taxonomy proposed by you, Ćirković and Bostrom) associated with human enhancement technologies?

Have we made significant progress in recent years, so that we would already be able to say, with some level of epistemic certainty, that a human enhancement technology would be high or low risk, so that such estimates would serve for the formulation of appropriate policies?

AS: I think that is an interesting issue. Generally, most human enhancements we are considering today, it's very hard to see how they could generate an existential risk, at least in a direct sense. If you think about cognitive enhancements, it's very unlikely that they could wipe out humanity. Now, on the other hand, if we start talking about genetic engineering, many people's intuition start to say: "wait a minute, that sounds like it could be more existential", because now we might also get intergenerational ethics. And I do think that what happens here is mostly that the means we might use to perform enhancement might have an existential risk aspect. The most obvious thing is that advanced biotechnology has tremendous ability to help us and harm us. So using genetic technology to construct new and dangerous pathogens is getting easier and easier all the time, so the technologies that would allow us to very easily modify our own biology would probably also allow a lot of agents to cause enormous harm, and that might lead to a situation where we could expect some form of biowarfare or bioterrorism. A world that has too much bio power that cannot be controlled might just be deadly. Similarly, other technologies that would be useful for generating enhancement like very advanced nanotechnology or very advanced artificial intelligence are inherently risky and potentially dangerous because they are powerful technologies and can change the world in unexpected ways, or malicious actors can use them in a way or we can get into conflict over it. But these are instrumental uses, so it's not so much the enhancement per se that has an effect. The Fukuyama theory is that you can get an existential disaster but we are willingly enhancing ourselves and losing something essential. We lose our potential and we might not recognize it. It's well worth noticing that Nick Bostrom has also made the same point. He's got some scary scenarios in one or two of his papers where he talks about the mindless outsourcers, one possible scenario where we gradually outsource the contents of our minds to more and more effective software, and eventually there's nobody home. There are no individuals. There is no consciousness. There is just this cloud of software maintaining an economy that is expanding and all the curves are going upwards forever, but there is nobody there. There is nobody to actually benefit from it. So that would represent a proper existential risk and it's an evolutionary pathway that we need to avoid. It's not so much that obvious enclosure disaster but we go down a pathway for values lost. A more recent reformulation of these ideas is found on Scott Alexander's brilliant essay *Meditations on Moloch*, where he actually combines Allen Ginsberg's beatnick poem *Moloch* with Nick Bostrom's work on superintelligence, and points out that we quite often have created these big systems, economies, institutions, rules that are intended to safeguard some values or help us get some values, but because of some bad design in the dynamics, they take over and they control the system. You can't opt out, because now you lose more than staying inside the system. But they also work against the values they were supposed to actually get. So the end result is that it's bad but you can't leave the game. That form of risk I think is really interesting because it is systemic. And it is much harder to analyze with the current academic tools. We understand asteroids much better than we understand the global supply chains. We understand even, I think, artificial intelligence better than we understand the global economy, which in many ways is like a cloud based super intelligence but actually solves allocation problems on a vast scale, produces products and focuses attention in a very complex way, which is very adaptive but not always aligned with our goals. So, getting to the question: are there other forms of enhancement that pose existential risk? I think there are enhancements that amplify certain kinds of risks. You could imagine a world where intelligence amplification made many actors smarter, which allowed them to do more things, and bad actors that could do worse. The question there is of course is whether the offense/defense balance remains the same. It could very well be, but a smarter world is a much safer world because it is actually quite hard to be a Professor Moriarty when there is a thousand Sherlock Holmes in the local police force. But you might also have other forms of amplification. For example, Persson and Savulescu in their book *Unfit for the*

Future responded a bit to the enhancement debate by suggesting that we need moral enhancement before we amplify our capabilities. In fact, they mirror in some sense Bostrom's and other work in the AI safety debate arguing that we need AI safety before we get powerful AI capability. So what Savulescu and Persson argued was that we need to enhance our moral capacities so that we can use our powers wisely. And then there are questions like if we can enhance empathy so on. But many of these enhancements might in themselves be risky. You can imagine a pill to strengthen group cohesion. You take the pill, you feel good with your group, and you are going to work well together. Does that mean that you are going to be nice to other groups? Unfortunately, many of the worst atrocities we have seen have been strongly connected groups working extremely well together against what they think are opposing groups. So just enhancing something doesn't necessarily mean it's going to be good. And you can get emergent problems. I'm somewhat optimistic, but even various attempts of moral enhancement might actually have surprising good effects. I have a paper where I looked at if you could change social value orientation, what would be the stable states. And I found to my great near coauthor and my own great surprise that actually prosocial orientation was the most stable one. In any case, I think there is option here for positive surprises, too. But generally, it is the systemic aspect of enhancement that is going to be the real challenge. This is where, I think, we are going to have to do a lot more work. It's going to be hard but worth it.

MK/MV: Professor Anders, we have one more question and we have one invitation to make. Would you have some extra time?

AS: Oh yeah, no problem.

MK/MV: So before our last question, we would like to commend your recent article published with your colleague Len Fisher ("A Safe Governance Space for Humanity"), in which you develop an interesting safe governance model for global catastrophic risks. Our question is based on two assumptions: (1) we have adequate anthropogenic and intermediary existential risk analysis methodologies to estimate possible harm; and (2) we have a good theoretical governance model capable of mitigating the harmful effects of a human enhancement technology. Even so, do you think that we have good reason to believe that the conditions for their effective implementation on a large scale will be present in the near future, so that we can continue to invest in the development of human enhancement technologies that imply significant risks (high or medium risks)? In other words, once the conceptual, epistemic and normative problems have been overcome (at the theoretical level), what is your expectation regarding the solution of what we will call the practical-political-economic problem of implementation?

AS: That was a wonderfully tricky question. I think it's a good one. So the paper I wrote with Len Fisher, it started when we were trying to figure out what could we actually say about governing complex adaptive systems and complex risks. We were a bit shocked at first when we realized that there are necessary conditions. Then we realized that these necessary conditions are actually… it's kind of obvious that they must exist. And there might actually be more necessary conditions but we didn't find, and so on. Many of them are very obvious because if you don't have them, you don't recognize that you're working inside a complex adaptive system, then your approaches are going to fail. But assuming that you actually understand the system fairly well, and that you have an idea about how to do it, that doesn't mean that you can easily implement it, because we have to work through actual human institutions with actual humans, actual economic interests, etc. And we end our paper on a kind of sneaky line that we suggest that the next question is: how do you implement this? The bridging conditions going from this one to one that actually implements it. And we're working on it. We're working on it. It is much harder. But one of the interesting realizations I had when dealing with the consequentialists and [effects-value of] the movement is to realize that Pareto improvements can sometimes be really worthwhile. Basically, if you make things better, that is on its own good even if it doesn't solve the whole problem. What you want to avoid is getting trapped in a trajectory that you cannot undo. You still want to keep things open. You still don't want to lock in things too hard. But generally, reducing existential risks, even slightly, is super valuable. Given how much is at stake in some moral sense, even a small reduction is worthwhile, so even a slight improvement in management of these risks is super valuable. Now the real question here is: can we say something about where path dependencies in the trajectories do show up? And unfortunately this is where things get complicated because we know human governance is one of the most path dependent things that exist. Partially, it is because it's by design. We create Constitutions to ensure that certain changes, once done, are very hard to undo. We create various legal frameworks that link backwards in time, that act as standards to stabilize things. Because those platforms may enable… they are good things to have. You can now build a society on top of that platform, and you don't want the foundations to be changing too fast. This, of course, is completely different from the adaptive governance we're talking about. And it also shows a limit to adaptive governance. You might sometimes want to reduce the capacity of a system simply because you might want to keep it adaptive or you might want to keep it rigid but easy to replace when it breaks. So I don't think there are any perfect solutions here. This is a very annoying point. I really wish I had this beautiful story about how if we just implemented this nice general scheme, it could be scaling endlessly to the limits of the universe. But fortunately we don't live in a simple universe where that would be possible. Sometimes we actually are going to have to make gambles on trajectories to select what seen promising given what we know but might lock us in later. A good example might be the United States that did some amazingly smart choices, Constitutionally speaking, back in the Revolution, and now we're stuck with a political system that is not really working well in the 21st century, but it is extremely harder for them to backtrack and change that. They might just have been unlucky or maybe that is what we should expect after a few hundred years for any Constitution. Similarly, when we start thinking about the policies for governing human enhancement, we can generally work towards things that are open ended that we can undo these issues, where new data allows us to change them, where we have institutions that are flexible and data driven, but some of them are going to be path dependent and "have lock in", and in that case we might just need to have multiplicity.

So my final point is an ecological analogy. In evolution and ecology, there is this concept of meta populations. Butterflies living in meadows [drearily] move between meadows because there are forests and other things in between, but sometimes they do. That means that each meadow has its own somewhat independent population. If the butterflies go extinct in one meadow, it gets recolonized eventually from others. This means that the overall system can become much more resilient than any individual part of the system. And I do think the same thing might be true also for the practical regulations. We want to ensure, in the face of a risk, that we have models that can fail without the cost being too unbearable. Then other models can reach in and help. And this can both be in terms of buildings. Modern fire codes in many places recognize that it is hard to prevent the contents of a room in case of a fire, but you can make sure the fire doesn't spread from the room to other parts of the building. We might think of it as ways of handling different countries. If one country implodes, other countries can reach out and try to help them. We might want to pick a multiplanetary species to reduce existential risk. And the really interesting question here is: what about enhancement techniques and regulatory systems? We might want to have different regimes. Normally, we tend to want to be universalist and say: shouldn't there be a moral principle that applies everywhere and to everyone? Yes, but we don't know which ones actually work well, especially together with a society. So we might actually want to have models, models that work differently, and if something fails we can learn from that and try to repair it. Again, this is not a perfect solution, but it's probably the best we can do, because our brains, enhanced or not, are very small compared to the universe. The universe is more complex than we are, and indeed what we can create, but we can learn, we can adapt, and I think that is the way of actually creating another activity that can survive indefinitely.

MK/MV: Professor Anders, it was amazing, fantastic. We are very honored to have this opportunity, and we are wondering if you could maybe next year come to Brazil…

AS: Oh, I would love to, yes!

MK: Murilo has some funds. Murilo works at Fiocruz – it's one of the most prominent institutes here in Brazil – and we are making some plans to have a workshop, a presential workshop here in Brazil in Rio where Murilo lives next year and we wonder if you could come to visit us and give us some time to give your lecture here. We have some community here of people working on this issue regarding existential risks, regarding your ideas. We have some good community, good people working on this stuff and so we would like to invite you next year maybe if you would like to come to Rio and well…

AS: I would be delighted to come. Yeah, we need to see of course if it works with schedule and stuff like that but generally yeah I would so would like to meet with you!

MK: Oh yeah, we appreciate, so Murilo then we are gonna keep in touch this idea and then we can send the date. I don't know, Murilo, if we have some schedule right now. I guess we don't, right?

MV: Not yet.

AS: No problem.

MV: Professor Sandberg, I would like to thank you deeply for this interview, for this significant academic contribution to advancing the debate and for taking the time to talk with us. We really hope we can meet face to face soon to continue this conversation to deepen the debate with you and learn more from you.

AS: Thank you!

MK: Professor Sandberg, professor Anders, Anders, my friend, we will keep in touch, we will send you this date for next year, for this presence, a presential workshop in Rio de Janeiro here in Brazil, and we have some funds. We are going to organize this and we will send you later, but first we would like to say again: thank you for this kindly interview. It was amazing. We have so much in common. We have some ideas that we will develop here in Brazil because some things that you said, they are really new and brand new and we have to make this community grow when it comes to risk, existential risk, human enhancement, artificial intelligence… So thank you again and we expect that we can meet each other next year here in Brazil. We are going continue this conversation, right?

AS: Wonderful! Yes see you next year in Brazil just keep in touch.

MK: Thank you professor, have a good day and a good afternoon, bye bye, see you.

Anders Sandberg: Bye bye see you.

MK/MV: Bye professor, thank you very much.

Submetido em 31 de maio de 2022.

Aceito em 25 de julho de 2022.