DELIBERATIO
STUDIES IN CONTEMPORARY PHILOSOPHICAL CHALLENGES

https://deliberatio.uvt.ro/

*1*(1) 2021, 77–86

ISSN 2810 – 5826 ISSN – L 2810 – 5532 Print
ISSN 2810 – 5532 ISSN – L 2810 – 5532 Online

# The World's Most Dangerous Idea? Transhumanism in the Age of Artificial Intelligence, Climate Change, and Existential Risk: Some Comments on Stefan Lorenz Sorgner's On Transhumanism

Sven Nyholm *

**Abstract**

In his book *On Transhumanism*, Stefan Lorenz Sorgner defends a particular version of transhumanism, which is inspired both by the English tradition of philosophy, including Mill's utilitarianism, and Nietzsche's philosophy. Sorgner rejects the idea of universally valid moral principles, and argues that we should be pluralists about what it might mean to live a good human life. Everyone ought to enjoy negative freedom of a form whereby they are not being hindered from using science and developments in modern technology in their quest to live a good life according to their own conceptions of the good. Nobody, however, ought to be forced to undergo a program of moral enhancement of the sort that Ingmar Persson and Julian Savulescu argue that we need in order to help us deal with the existential risks we are facing, which our evolved human psychology has not prepared us to deal with. In this commentary, I argue that while Sorgner's techno-optimistic outlook is inspiring, we should not ignore issues related to existential risks created by modern technologies of the sorts that Persson and Savulescu discuss. Artificial intelligence and other developing technologies pose great risks, including that of contributing to climate change, and so transhumanists should concern themselves with the existential risks related to modern technology, and not only the potential benefits of these technologies. In addition to articulating that point, I also discuss Sorgner's metaethical assumptions, his claims about value, and the comparison between Nietzsche and Nick Bostrom that Sorgner makes in the book.

*Keywords:* transhumanism, artificial intelligence, existential risks, climate change, values

## 1. Introduction

In philosophy, we have this strange practice of paying tribute to each other by criticizing each other's work. Since I enjoyed reading and thinking about Stefan Lorenz Sorgner's (2016/2020) book *On Transhumanism*, I am glad to have this opportunity to engage critically with it, and to thereby pay tribute to Sorgner and his work. My remarks in what follows are based on a presentation I gave at a symposium organized around the book in May of 2021, where I and others took turns at criticizing different aspects of the book, and Sorgner patiently and enthusiastically responded to each of those commentaries.[1] It was nice to discuss Sorgner's

[1] Video recordings of the symposium are available here (part 1): https://youtu.be/rtvIMLMP-BM and here (part 2): https://youtu.be/mJAZHmCGuYY (accessed on October 1, 2021)

* *Department of Philosophy and Religious Studies, Utrecht University, Janskerkhof 13, 3512 BL, The Netherlands. E-mail: s.r.nyholm@uu.nl; https://orcid.org/0000-0002-3836-5932.*

ideas with him on that occasion, and it is nice to have the opportunity to continue the discussion here.

Notably, *On Transhumanism* covers a lot of ground. The discussion is conducted at a wide range of philosophical levels of abstraction. It contains everything from metaethics to normative ethics, from applied ethics to political theory, metaphysics, ontology, the history of philosophy, and many more things. In my engagement with the book below, I am going to be jumping around a little. I will comment on different aspects of Sorgner's discussion in the book, touching on topics on different levels of abstraction, relating to metaethics, normative ethics, and more practical, applied ethics issues respectively.

First, I will argue that somebody who is as techno-optimistic as Sorgner is, should—or, at least, it would be interesting to hear him—say more about the existential risks that are related to artificial intelligence and other powerful modern technologies. After that, I will briefly discuss Sorgner's claims about values and meta-ethics, and then end by very briefly discussing the relation between Friedrich Nietzsche (as portrayed by Sorgner) and Nick Bostrom (again, as he is portrayed by Sorgner).

## 2. The Most Dangerous Idea in The World?

Transhumanism—which, not surprisingly, is what *On Transhumanism* is about—is either a philosophical idea/theory or a movement, or perhaps both, depending on how you look at it. Whether it is regarded in its more academic guise, as a set of ideas and arguments defending those ideas, or in its guise as a movement, where people have formed political parties and done other things to spread and implement these ideas in society—there are some key things that all forms of transhumanism have in common, even if this is a rather "broad church" and many different types of ideas and theories can be found under the umbrella of transhumanism.

One thing that all of the different types of transhumanism have in common is the idea that our current ways of being or functioning as humans—our capacities, our nature, and so on—are regarded as not being optimal, but as being worth improving upon (Helmus, 2020). Another thing that I understand all different forms of transhumanism as having in common is the idea that we can use modern science and technologies to change and improve—to enhance—human beings. In short, we are not in an optimal state, there is room for improvement, and we should use science and technology to enhance ourselves. Different transhumanists differ in terms of what they would like to improve about human beings, what means they think could be used, and how radical they are in terms of how much we need or should enhance ourselves (More & Vita-More, 2013; Sorgner, 2016/2020).

The basic idea or set of ideas briefly sketched above were referred to as "the most dangerous idea in the world" in a 2004 article in *Foreign Policy*, in which Francis Fukuyama criticizes transhumanism (More & Vita-More, 2013, part IX; Fukuyama, 2006). Not surprisingly, Sorgner discusses that assessment from Fukuyama in his book. Accordingly, one of the things one cannot help but to think about as one is reading *On Transhumanism* is whether what Sorgner is presenting—his version of transhumanism—is something that could be thought of as being the most dangerous idea in the world. As it happens, there were some passages where this thought came to mind because it seemed to me that what Sorgner was presenting does not sound very dangerous or controversial at all. For example, Sorgner (2016/2020) writes:

> . . . I do believe that constant self-overcoming is central to promoting my own quality of life. I also consider scientific research, especially in biotechnology, extremely important and advocate for greater sponsorship of those research fields. I consider the availability of anesthetics, vaccinations, and antibiotics important achievements. I hope that future achievements will also address important challenges. This stance can be parsed as a weak form of transhumanism. (p. 14)

What Sorgner here and elsewhere in the book calls a "weak form of transhumanism" strikes me as what one might also call a fairly "mild" form of transhumanism. Sometimes the view that Sorgner presents even strikes me as not being very much like what one typically associates with transhumanism at all, but rather a big dose of technology optimism and great enthusiasm about the potential for science to improve our lives—something that even somebody like Fukuyama could potentially agree with. Well, perhaps Fukuyama would not be willing to say that he agrees with any form of transhumanism, whether it is "weak" or "mild" (Fukuyama, 2006). But the ideas that Sorgner presents in the quote above as being at the heart of his transhumanism seem like ideas that very many people would accept, whether or not they would call themselves transhumanists. I am not sure whether I would call myself a transhumanist, for example. But I happily agree with Sorgner that constant self-overcoming is likely to be central to improving the quality of life for many, if not most, people; I endorse scientific research, including biotechnology, and am also in favor of greater sponsorship of those research fields; and what Sorgner lists as great achievements above are certainly things I would follow him in celebrating as great achievements. I am just not sure that following Sorgner's lead on those sorts of points qualifies as something one should label transhumanism. Or, alternatively put, perhaps very many of us in modern, largely secular societies are transhumanists, at least of the "weak" or "mild" form, even if not all of us use that term about ourselves.

There is another part of Sorgner's view that could perhaps be seen as more dangerous—namely, a part of Sorgner's view that is much more conventional than the views of some more radical transhumanist philosophers and defenders of human enhancement. I am referring in particular to Sorgner's rejection of the idea that what Ingmar Persson and Julian Savulescu (2012) call "biomedical moral enhancement" would be a good idea or something transhumanists should argue in favor of.

According to Persson and Savulescu, we need moral enhancement—that is, enhancement of our capacities for moral emotions and moral behavior—because our evolved human psychology is better adapted for life in small, pre-modern societies than life in the modern world, with huge nation states, collective action problems leading to climate change and other existential risks, and modern technologies that we lack proper control over. Our emotional capacities are well-adapted to deal with interpersonal conflicts on a small scale—a type of situation human beings repeatedly faced during the long period of human evolution. But our emotional capacities are not well-adapted to deal with large scale problems related to life in the modern world, again, such as human-caused climate change. We are, Persson and Savulescu argue, "unfit for the future". For this reason, we need to develop technologies that can be used for moral enhancement of us human beings. Or so Persson and Savulescu argue. (Persson & Savulescu, 2012, Chapter 1) Sorgner, however, takes a skeptical stance towards this idea, just like many authors who have somewhat more conventional ideas about human morality do. Sorgner (2016/2020) writes:

> The background against which ethicists like Savulescu address the question of moral enhancement is the view that the global problems that could result in the extinction of humanity must be solved fast. Does this not mean, however, that these technologies must then also be used globally against the will of individuals? (p. 29)

Sorgner's (2016/2020) response to Persson and Savulescu's argument continues as follows:

> Such a measure would neither be easy to implement nor desirable, as it would undermine the cherished and fragile achievement of the norm of negative freedom that allows people to determine their own actions according to their own idiosyncratic notion of what they consider living well. I advocate strengthening this standard even further rather than restricting its effectiveness through global, paternalistic, coercive measures. Undermining this norm indicates totalitarian tendencies. Historical experiences with the Third Reich and other dictatorial regimes should have made it clear to us, however, that such movements must be avoided at all costs . . . . (p. 29)

Sorgner (2016/2020) concludes: ". . . reflecting on moral enhancement technologies seems less relevant to me than reflecting on other types of enhancement meant to promote human performance" (p. 30).

In sum, Sorgner rejects Persson and Savulescu's idea about a need for moral enhancement because it is or appears to be paternalistic, coercive, and reminiscent of totalitarianism. But what about the existential risk worries that Persson and Savulescu talk about when they argue in favor of the need for moral enhancement?

I develop a similar line of argument related to robots and artificial intelligence (AI) in my book *Humans and Robots: Ethics, Agency, and Anthropomorphism* (Nyholm, 2020). In the first chapter of that book, I ask: with our human psychology based on human evolution, are we ready for a world increasingly filled with robots and advanced AI? People tend, for example, to respond to robots and AI systems in anthropomorphizing ways—i.e., as if they have human qualities—and they tend to be misled in various ways by the modern technologies we use. Most people do not understand how advanced modern technologies work, and tend to underestimate their capacities in certain ways while also overestimating their capacities in other ways. So, either we have to take many kinds of precautions that we are currently not taking in developing safe robot and AI technologies, or we need to try to make ourselves better at interacting with the robots and AI technologies we are developing. Business as usual may be highly unsustainable. Simply developing advanced robots and AI systems without thinking about whether our human psychology is well-adapted to deal with them creates or might create serious risks—including existential risks—for human beings (Nyholm, 2020, Chapter 1; Cf. Ord, 2020).

Sorgner's techno-optimism seemingly leads him to ignore or at least downplay possible risks and threats posed by technologies and life in the modern world of the sort that Persson and Savulescu describe, as well as the sorts of risks related to advanced robots and AI of the sorts that I and others are concerned about. Sorgner is against moral enhancement. But is he also against worrying about existential risks? Notably, Sorgner follows Persson and Savulescu in taking an evolutionary view of human beings. But does he reject their view that human evolution has made us poorly prepared for life in modern society, with modern technologies? Not addressing such worries, i.e., not addressing whether we are "fit" or "unfit for the future" and potentially in need of some sort of moral enhancement, while at the same time being very enthusiastic about the positive potential of emerging technologies—that might be the most dangerous idea in the world.

It should be noted here that existential risks from something like AI might not need to come from the sort of super-intelligent AI systems that Nick Bostrom (2014) describes in his fascinating book *Superintelligence* (cf. Häggström, 2021). Rather, much less intelligent AI systems than super-intelligent ones might pose great threats to human beings for other reasons than being super-intelligent beings that might try to "take over". For example, a nuclear war almost started in 1983 when a Soviet Union nuclear attack detection system falsely indicated that the United States had launched missiles that, according to this warning system, were supposedly heading towards the Soviet Union. The Soviet protocol required the officer at the command center to directly report the messages from the detection system to the higher-ups on the chain of command, so that they could prepare a retaliatory attack. However, the officer on duty, Stanislav Petrov, potentially "saved the world" by breaking with protocol, because he judged that the message from the system was most likely a false alarm. This was a case where the use of a not very intelligent—and certainly not super-intelligent—basic AI system might have led to large-scale damage in the form of an international nuclear war (for discussion of this case, see Nyholm, 2021).

Another kind of threat that contemporary AI systems pose is also that they have an enormous carbon footprint—as Kate Crawford (2021), Aimee van Wynsberghe (2021), Mark Coeckelbergh (2021), and others note. Without being super-intelligent, then, AI systems can also pose

existential threats by contributing greatly to the climate change threat. In our enthusiasm about developing modern technologies, in other words, we might create technological systems that drain the world's resources, cause pollution, and in other ways wreck the climate.

In mentioning these things, I do not mean to suggest that Sorgner is giving evidence of being reckless in his technology optimism. Nor do I mean to disagree with any of the worries about totalitarianism and so on that Sorgner expresses about large-scale, coercively enforced moral enhancement. Those are all very valid concerns. What I would be interested in hearing about from Sorgner's side is his take on the kinds of problems that Persson and Savulescu describe that lead them to argue that there is a need for moral enhancement, as well as his take on the kinds of existential risks that are posed by advanced modern technologies—in particular, AI—and whether our evolved human psychology is well-adapted to deal with these problems. In other words, Sorgner argues against the conclusion of arguments such as the one offered by Persson and Savulescu. But he has less to say about the steps of the argument leading up towards the conclusion that we need some form of moral enhancement. At least, that was something that was missing from *On Transhumanism*, but which Sorgner might potentially go on to address in work he will be doing in the future. I, for one, would be very interested in his take on these issues.

## 3. Objective Values and Changes in Values

Having discussed the more practical or applied issues above, I will now turn to a more general level of philosophical abstraction and discuss Sorgner's meta-ethical views about value and value-change. In his meta-ethical theorizing, Sorgner follows Nietzsche's lead. Or rather, it is perhaps better to say that he follows one possible interpretation of Nietzsche, since Nietzsche is a philosopher of whose work there can be many different interpretations. Commenting on Nietzsche in a way that endorses his view, Sorgner (2016/2020) writes that on the sort of view he and Nietzsche share:

> Values are apt to change on cultural, social, and personal levels. . . . There are no absolute and unchanging values since he rejects the existence of the Platonic world of ideas that would be the necessary foundation for any such permanent values. (pp. 60-61)

In other words, Sorgner rejects the idea that we can justifiably speak about universal or objective values in any sensible way. Considering the perspective of those who think there can be generally valid views about what is a good human life, on the one hand, Sorgner (2016/2020) writes:

> I, on the other hand, believe in a radical plurality of the good. Any attempt at a nonformal determination of the good seems to me to be condemned to failure since individual physiopsychological requirements differ too greatly to allow for any nonformal, universally valid description of the good life. (p. 84)

My response to the view laid out in these quotes is that it may involve a false dichotomy. Is it really the case, I want to ask, that we should think that either values are in constant flux or they are permanent Platonic forms? And does rejecting moral realism of a Platonic kind really require us to reject any philosophizing and theorizing about substantial moral ideals and values, which we might invite others to join us in endorsing? It seems to me that we should not think that either one endorses a view of values as being in constant flux and as involving a radical plurality of the good or one needs to be a Platonist about values and regard them as permanent Platonic forms. These are, rather, the extreme ends of a spectrum of possible meta-ethical views. And views that are not at these extreme ends of the spectrum of possible meta-ethical views are more plausible

than the views on either extreme end of the spectrum between hard-core relativism and hard-core realism.

There has been some convergence in recent meta-ethics around the view that it is possible to make claims about substantial values or endorse what are treated as generally valid moral principles without thereby committing oneself to a Platonic form of moral realism. On the one hand, there are so-called expressivist theories, such as those defended by Allan Gibbard (2003) and Simon Blackburn (1998), on which expressing our acceptance of norms that are viewed as universally valid does not commit one to anything metaphysical or ontological, but on which it rather involves a form of strong practical commitment to the norms or values that one endorses. On the other hand, there are recent forms of moral realism—such as those defended by Thomas Nagel (1997), T.M. Scanlon (2013), and Derek Parfit (2017)—on which being committed to the idea that there are objective/non-subjective values and generally valid moral principles also does not involve any metaphysical or ontological commitments. According to Ronald Dworkin (2011), for example, what we are committed to in treating something as an "objective" value or a generally valid moral principle is that we are willing to argue in favor of these values and to defend them against skeptical challenges. According to Scanlon (2013), in turn, so long as some domain of human life has its own internal standards regarding how to argue and reason about the subject matter of that domain of life, there is no reason to take a skeptical or relativistic view about that subject matter. Ethics is one such part of life, where there are types of reasoning and arguments that do not reduce to the types of reasoning and arguments that we use in other domains (such as natural science, legal reasoning, or mathematics). Accordingly, we can treat ethics as a domain with its own standards, Scanlon argues, and there is no reason to take a skeptical view about ethical reasoning just because one wants to reject extreme Platonic versions of moral realism.

There are also naturalist versions of moral realism, such as the type of theory that is defended by Peter Railton (2017). According to Parfit (2017), in the main line of argument laid out in his book *On What Matters: Volume Three*, there is a great deal of convergence among the expressivist, realist, and naturalist realist views that he and the above-mentioned authors have developed and that others have also participated in developing. The convergence is precisely around the idea that one does not need to be a Platonic moral realist, who postulates metaphysically dubious permanent forms, in order to be able to vindicate ethical reasoning involving substantive values and ethical principles. Whether Parfit and these other authors are right about this is not something that can be settled here or that I am able to discuss further in this paper, but the fact of this emerging agreement about these things among these influential philosophers provides some reason for thinking that Sorgner may be presenting us with a false dichotomy—or at least an exaggerated dichotomy—in the first quote above.

Regarding the second quote above, it is not altogether clear to me what Sorgner has in mind when he speaks about a "universally valid description of the good life" and how it is impossible because people differ from each other. If Sorgner means that there can be no universal valid description of the good life—not even in general terms—because people tend to disagree about what is most important in life, then I think we should reject the underlying premise in Sorgner's reasoning. Normative concepts have as one of their key features, as I see things, that we can use them to state disagreements about what is good, right, just, or whatever. This point can be illustrated with an example. Imagine that I say the following: "I think that self-improvement, mutual care and respect, and the promotion of human knowledge are essential parts of the good life, but there are those who disagree with me." Such a statement does not involve any kind of contradiction. Instead, it recognizes that there is disagreement while expressing a substantive commitment to a certain conception—or partial conception—of the good life (Blackburn, 1998; Gibbard, 2003). Nor is the validity of a certain conception of the good life automatically undermined by the fact that some people may disagree with it. Perhaps the fact that others

disagree with us should make us humble and careful about placing too much trust in our own views. But it is not, all by itself, a reason to think that neither side could be right and that we have to view our own conceptions of the good life in relativist terms (Dworkin, 2011).

It may be that I am reading too much into the quotes above and that Sorgner is not meaning to rule out the possibility of substantive ethical arguments. And it may be that he is also simply urging us to be open-minded about other ethical perspectives and to exercise intellectual humility about our own. After all, Sorgner himself enthusiastically endorses the value of negative freedom, and he does this in a way that celebrates that value without claiming to be able to with certainty prove its universal validity. But if Sorgner can put forward one substantive value or ethical principle in his arguments—namely, the promotion of negative freedom—then why could he not also allow for other values of "non-formal" kinds, assuming that the value of negative freedom is not a purely formal one? If Sorgner can celebrate the norm of negative freedom as a "great achievement" without committing himself to Platonic forms, why could he—or others—not also appeal to other substantive values or ethical principles and treat them as generally valid without committing to Platonic forms or other metaphysically extravagant ideas? That was a question that I was left with after reading Sorgner's discussion of values.

## 4. Nietzsche and (or versus) Bostrom

Speaking of Sorgner's meta-ethical views and his more general views about values, I will end with some more picky, if not nit-picky, commentary. Specifically, I will make some very brief comments about Nietzsche (on whose philosophy I am admittedly no expert!) and then also some brief, but not quite as brief remarks about Bostrom and what Sorgner says about his views about value. Let us start with Nietzsche and whether he can be seen as a transhumanist of the sort that Sorgner portrays him as being. In particular, let us very briefly consider whether there is not more of a clash between Nietzsche's philosophy and what Sorgner presents as the roots of transhumanism than Sorgner thinks. At one point, Sorgner (2016/2020) writes: "Transhumanists are deeply rooted in the English tradition, which is closely linked to Darwin's theory of evolution and Mill's utilitarianism" (p. 40).

My very brief comment is simply that this claim about the deep roots of transhumanism seems to place Nietzsche's philosophy in stark contrast with transhumanist philosophy, since Nietzsche—at least in my non-expert reading—was deeply skeptical of most English philosophy. Was Nietzsche not both an opponent of Darwin's theory of evolution or key parts of it and, even more so, an opponent of Mill's utilitarianism? If so, it might seem as a bit of a stretch to develop a Nietzschean form of transhumanism if we at the same time want to understand transhumanism as being deeply rooted in the English tradition with a focus on Darwin's theory of evolution and Mill's utilitarianism (cf. Hauskeller, 2010). But, like I said, I do not claim expertise in Nietzsche exegesis, and Sorgner is an expert on Nietzsche's philosophy, so perhaps these brief remarks should be primarily seen as being articulations of the sort of confusion that a Nietzsche exegesis novice finds him or herself with when reflecting on the idea of a Nietzsche-inspired form of transhumanism.

Speaking of Nietzsche, Sorgner also compares Nietzsche's views with those of Bostrom. Here, too, I find myself a little bit puzzled. Sorgner (2016/2020) writes the following about Nietzsche and Bostrom: "Besides the ontological dynamism that can be found both in transhumanism and in Nietzsche's philosophy, they also both exhibit similar flexibility with regard to values" (p. 60).

To substantiate this further, Sorgner quotes the following passage from Bostrom (2005): "Transhumanism is a dynamic philosophy, intended to evolve as new information becomes available or challenges emerge. One transhumanist value is therefore to cultivate a questioning

attitude and a willingness to revise one's beliefs and assumptions" (as cited in Sorgner, 2016/2020, p. 60).

Commenting on that quote, Sorgner (2016/2020) writes that "Nietzsche makes a similar point to Bostrom's", namely, "that prevailing values have constantly changed." (p. 60). Sorgner also points to the following Bostrom (2005) quote: "Transhumanists insist that our received moral precepts and intuitions are not in general sufficient to guide policy" (as cited in Sorgner, 2016/2020, p. 61). As Sorgner reads Bostrom, Bostrom uses that idea to then go on to propose a set of values that involve a "dynamic worldview". In particular, Sorgner also highlights the following quote from Bostrom (2005):

> We can thus include in our list of transhumanist values that of promoting understanding of where we are and where we are headed. This value encloses others: critical thinking, open-mindedness, scientific inquiry, and open discussion are all important helps for increasing society's intellectual readiness. (as cited in Sorgner, 2016/2020, p. 61)

Commenting on that quote, Sorgner (2016/2020) writes that "Nietzsche largely concurs with this statement" (p. 61). I do not wish to dispute that Nietzsche would endorse the values of critical thinking, open-mindedness, scientific inquiry, and open discussion. Nor would I dispute that Nietzsche would endorse the value of promoting understanding of where we are and where we are headed. What I am more skeptical about is whether what we see in the Bostrom passages quoted above—and in his view more generally—is an example of "flexibility with regard to values" in the sense that Bostrom has a view on which the correct values, so to speak, are in a dynamic flux.

As I read him, Bostrom does not seem to be saying that transhumanist values are in constant flux and need of continual updating. Bostrom rather seems to firmly assert or endorse certain values such as a commitment to critical thinking, science, self-improvement, and so on. We may need to update our scientific view of the world and understanding of ourselves as science progresses and new technologies become available. But thinking that is perfectly compatible with being a little more rigid in one's adherence to certain basic transhumanist values and principles, which Bostrom seems to be his papers in which he is defending transhumanism. His adherence to transhumanist values does not seem to be something that is presented as a form of flux or dynamically changing process. In other words, Bostrom seems to me to be more of a "value realist", if you will, than Nietzsche, who seems to be more of a "value relativist", or, in Sorgner's preferred terminology, a "value perspectivalist" (Sorgner, 2016/2020, p. 64). There may be some overlap in the views of Bostrom and Nietzsche in some regards, to be sure, but I suspect that the divergence between their respective views is larger than Sorgner makes it out to be.

I will make one last observation about Bostrom, as his philosophy features in *On Transhumanism*, before I end my commentary. And that observation is that Sorgner seems to focus completely on what might be called the "early Bostrom" and that he does not say anything about what we might call the "later" or, perhaps, "more mature Bostrom". The early Bostrom—at least from a zoomed-out perspective—comes across as exuding strong techno-optimism with various essays passionately defending transhumanism. More recently, the later or perhaps more mature Bostrom—again, at least from a zoomed-out perspective—comes across as being much more worried about technological risks, specifically existential risks related to superintelligence and other forms of advanced AI.

When I made this observation at the above-mentioned symposium organized around Sorgner's *On Transhumanism*, David Wood, of the London Futurists, was in the audience. He commented that there is techno-optimism both in Bostrom's earlier work and in his later work, as well as some degree of techno-pessimism in both Bostrom's earlier and more recent work. That is

certainly true. That is, it is certainly true that Bostrom did not switch from being a complete techno-optimist to being a complete techno-pessimist. However, on the whole, the themes from Bostrom's earlier work that made the biggest splash were the techno-optimistic aspects of those earlier essays. Moreover, the themes from Bostrom's more recent work that have made the biggest impact on practical philosophy and AI research are those that have to do with the existential risks and potential dangers of future super-intelligent AI technologies. And while Sorgner frequently refers to Bostrom's earlier techno-optimistic ideas about transhumanism in *On Transhumanism*, he does not engage with Bostrom's later work about existential risks related to future AI presented in Bostrom's *Superintelligence* in the book. It would, however, have been very interesting to hear how Sorgner would try to balance his techno-optimistic ideas that are in line with Bostrom's earlier work against the worries about emerging technologies that are now also prominent in Bostrom's work, in particular his more recent work.

In other words, I end this section on my commentary like I ended the first main section above, i.e., with a plea to Sorgner that he should devote more of his thinking to potential risks—including so-called existential risks—related to modern technologies, and perhaps particularly in relation to developing and potential future forms of artificial intelligence. By making this plea, I do not mean to suggest that Sorgner should change the tenor of his overall techno-optimism and adopt a more pessimistic or cynical outlook on modern technology and its potential. I do not mean to suggest this, because I find Sorgner's enthusiasm about modern technologies and the possible future of humanity to be an inspiring and a worthy contribution to the debate about what path we should take as we move into the future. But it would be very interesting to hear from Sorgner what he would say about the kinds of risks and indeed existential risks that are also part of what we need to deal with as we continually develop increasingly powerful technologies—including more advanced forms of AI—that might potentially slip out of our control.

## 5. Concluding Remarks

As I mentioned in my opening remarks above, philosophers have the odd habit of paying tribute to each other by criticizing each other's work. I have used this opportunity to do just that, and was glad to be able to do so since I very much enjoyed reading and critically engaging with Sorgner's book *On Transhumanism*. The above-mentioned symposium where I and others offered critical responses to Sorgner's book was not only intellectually stimulating, but also a testament to Sorgner's willingness to engage with criticisms of his work in an enthusiastic, mutually respectful, and constructive way. I hope—or, rather, I am sure—that my critical engagement with Sorgner's book in this written version will also be met with the same openness and willingness to reflect together on these ideas, just as we did on the occasion of the symposium.

## Acknowledgements

My thanks to Aura Schussler, David Wood, Karim Jebari and, in particular, Stefan Sorgner.

## Funding

My work on this paper is part of the research program "Ethics of Socially Disruptive Technologies", which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

## References


Blackburn, S. (1998). *Ruling passions*. Oxford University Press.

Bostrom, N. (2005). Transhumanist values. *Journal of Philosophical Research*, *30* (Supplement), 3–14.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Coeckelbergh, M. (2021). *Green Leviathan or the poetics of political liberty: Navigating freedom in the age of climate change and artificial intelligence.* Routledge.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dworkin, R. (2011). *Justice for hedgehogs*. Harvard University Press.

Gibbard, A. (2003). *Thinking how to live*. Harvard University Press.

Häggström, O. (2021). *Tänkande Maskiner*. Fri Tanke.

Hauskeller, M. (2010). Nietzsche, the overhuman and the posthuman: A reply to Stefan Sorgner. *Journal of Evolution and Technology*, *21*(1), 5–8.

Helmus, C. (2020). *Transhumanismus—der neue (Unter-)Gang des Menschen? Das Menschenbild des Transhumanismus und seine Herausforderung für die Theologische Anthropologie*. Verlag Friedrich Pustet.

More, M. & N. V. More. (2013). *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future.* Wiley-Blackwell.

Nagel, T. (1997). *The last word*. Oxford University Press.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield International.

Nyholm, S. (2021). Meaning and anti-meaning in life and what happens after we die. *Royal Institute of Philosophy, Supplements* 90: in press.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hatchett Books.

Parfit, D. (2017). *On what matters* (Vol. 3). Oxford University Press.

Persson, I. & Savuelscu, J. (2012). *Unfit for the future*. Oxford University Press.

Railton, P. (2017). Two sides of the meta-ethical mountain?, In Peter Singer (Ed.), *Does anything really matter? Essays on Parfit on objectivity* (pp. 35–60). Oxford University Press.

Scanlon, T.M. (2013). *Being realistic about reasons*. Oxford University Press.

Sorgner, S. L. (2020). *On Transhumanism: The most dangerous idea in the world?!* (S. Hawkins, Trans). Pennsylvania State University Press. (Original work published 2016).

Van Wynsberghe, A. (2021). Sustainable AI: AI *for* sustainability and the sustainability of AI. *AI and Ethics 1*(3), 213–218.