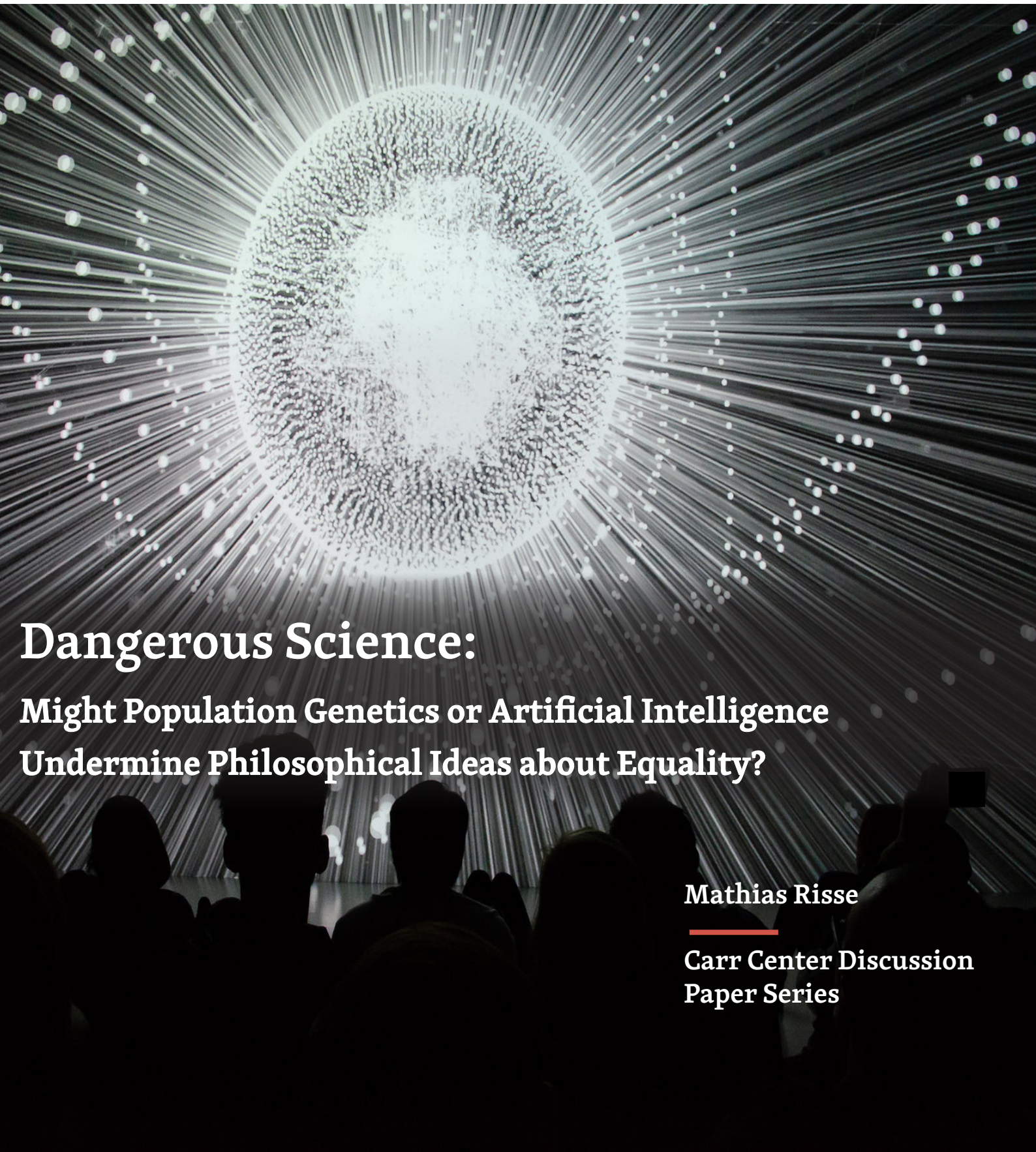
HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

SPRING 2020
ISSUE 2020 - 010

CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL



Dangerous Science:

Might Population Genetics or Artificial Intelligence Undermine Philosophical Ideas about Equality?

Mathias Risse

Carr Center Discussion Paper Series

Dangerous Science:
Might Population Genetics or Artificial Intelligence Undermine Philosophical Ideas about Equality?

Mathias Risse

Lucius N. Littauer Professor of Philosophy and Public Administration
Director of the Carr Center for Human Rights Policy
Harvard University

July, 2020

This paper was prepared for an interdisciplinary conference on *Gefährliche Forschung?* (Dangerous Science?) held at the University of Cologne in February 2020 and is scheduled to appear in a volume of contributions from that event edited by Wilfried Hinsch and Susanne Brandstätter (the organizers) and to be published by de Gruyter. The question posed to me was precisely the question in the title, so the paper goes right into responding to that question without locating the context of this debate or offering a preview of its contents. The first section discusses the ideal of equality, the next two talk about genetics in the context of responses to racism, and the remaining two speak about possible changes that might come from the development of general Artificial Intelligence. I'm grateful to the organizers and other participants for discussion, to Sushma Raman for helpful comments and to Derya Honca for editorial assistance.

Table of Contents

1. Equality

In one way of reading him, Thomas Hobbes – one of the founding spirits of modern political philosophy in the 17th century - thought it was a natural-law requirement that people recognize each other as equal although they are not. It is necessary for peace and the maintenance of society to do so. Only then can people escape from disastrous confrontations resulting from pride, contempt, and open disagreement about comparative worth. As Hobbes explains in his ***Elements of Law***, equality "by nature" is not descriptive but a principle that "men considered in mere nature ought to admit among themselves," or, as he says in framing this conclusion as a natural law: "Consequently, we are to suppose, that for peace sake, nature hath ordained this law. That every man acknowledge other for his equal. And the breach of this law, is that we call PRIDE."[1] So Hobbes' natural men are neither "born" nor "created" equal. Still, they have good reason to mutually acknowledge each other as equals: that is the only way to fend off unending conflict about who is superior to whom and what that would entail.[2]

Such an understanding of equality is rather sensible, regardless of whether it is Hobbes we have to thank for it. To be sure, a natural law of equality for the sake of peace is not what people normally have in mind when talking about "equality" among humans. I set aside theistic contexts, where "equality" tends to be something humans have in common in their relationship to God, which Christianity develops in terms of us being children of God. Outside of such contexts, talk of equality among humans normally means either "moral equality" or "equality among citizens." "Moral equality" (or the "equal worth" of all human life) concerns the distinctively human life, characterized at the species-level by a capacity for conscious reflection that involves an at least subjective awareness that we can make choices, and by a strongly developed capacity to cooperate. Moral equality could be understood naturalistically: it is the kind of brain we have (with its complexity and power), plus the kind of educational investments family and society normally make in us, that render each person worthy of certain expressions of respect, and of certain measures of protection and support.[3] The 1948 Universal Declaration of Human Rights (UDHR) is one prominent proposal for what such moral equality should involve practically.

The second common meaning of equality is equality among citizens. In modern states, people participate in the maintenance of intensely cooperative orders where compliance is enforced through the constant possibility of coercion. The way the cooperative order is arranged affects us rather profoundly, shaping who we get to be to begin with, and involves a continuous willingness to subject oneself to societal expectations. We all find ourselves in this highly intrusive situation. In return can expect to be treated as equal participants in the design of a system of political and economic rules that, after all, is largely conventional.

We can each expect that the many ways in which society makes rules (including rules of property) give our interests due consideration. This kind of equality could be spelled out in terms of John Rawls' two principles of justice. The first guarantees to each citizen the maximum level of civil and political rights compatible with the same rights for everybody else. The second assures everybody of genuinely fair equality of opportunity in the education sector and with regard to other ways that lead to some kind of privilege, and in addition judges only those economic inequalities as acceptable that are to everybody's advantage.[4]

Both understandings (moral equality and equality among citizens) leave much room for legitimate inequalities, especially material ones, as long as core tenets of those two standpoints are satisfied. But both times, the way the basic egalitarian thought is grounded is also open to doubt. What about all those obvious inequalities that seem to matter in day-to-day life, regarding intelligence, strength, attractiveness, empathy, ability to get along, etc.? Do not these make us different both one-human-to-another, and as cooperators in society? There are two response strategies. One is to insist the differences pale compared to what we share at the species level. As philosophers would say, the relevant equality is a ***range property*** (as in, you are either in the circle or not); what matters is to be inside, but if you are, you might be closer to or further from its center. The other strategy we can take from Hobbes. Moral equality and equality among citizens provide prima facie rationales of why people would be equal. The Hobbesian move could quell doubts about whether certain inequalities should not trump the

[1] Hobbes, *Human Nature and De Corpore Politico*, 78, 93.

[2] This (controversial) interpretation of Hobbes draws on Hoekstra, "Hobbesian Equality."

[3] This notion of equality is explicitly developed in the work of Ronald Dworkin, see e.g., Dworkin, *Life's Dominion*; Dworkin, *Sovreign Virtue*. Any view that spells out what human equality is based on rather than simply stipulating equal worth of human life has to meet conceptual challenges of finding room for disability; see e.g., Nussbaum, *Frontiers of Justice*, chapters 2-3.

[4] Rawls, *A Theory of Justice*; Rawls, *Restatement*. For these notions of equality see also Risse, *On Global Justice*, Part I.

equality-grounding considerations after all. In other words, that move responds to those who insist that some ways in which we are unequal should matter more than the ways in which we are equal and that matter prominently in the formulation of moral equality and equality among citizens.[5]

2. Race

Ever since humans have lived together in societies, they seem to find ways to introduce hierarchies, ways of putting some above others. For many generations, it would have been inconceivable to see themselves as living lives of equals. The UDHR is a truly historical document because it not only spells out what human equality involves but does so by providing a moral blueprint for a network of organizations designed to have global reach (the United Nations and the organizations in its system). But its framers were so fundamentally worried about the reassertion of hierarchical societies that they endorsed non-discrimination in different ways in the first two articles:[6]

> Article 1:
>
> *All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.*
>
> Article 2:
>
> *Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.*

That is, Article 1 states the basic point about equality. Article 2 makes sure we do not make exceptions by excluding certain groups from the scope of equality. One typical ground on which such exception would be made, especially in the decades preceding the UDHR, is race, listed in Article 2 as the first ground on which no exceptions ought to be made.[7]

The framers would have thought of National Socialism's racist ideologies that recently had sent millions to their deaths. However, as they were aware, racism was common in the world European colonialism had created since the late 15th century. Racial naturalism provided much of the glue for the white supremacism that dominated this period. Racial naturalism depicts a set of human races as bearing bio-behavioral essences: there are underlying natural properties that (1) are heritable, (2) are shared by all and only members of a race, and (3) explain behavior, character and culture.[8] As an example of how perniciously and lastingly this doctrine worked, consider how slavery, one way or another, has to some extent shaped the majority of countries in the Americas. The enslavement of Africans would have been impossible to justify other than through such a doctrine.[9]

The UDHR testified to the fact that racial naturalism would not be the credo for the political and economic system built after the Second World War. But then what to make of "races"? Is there still any grounding for talk about this kind of distinctions? Two major responses have long been available. One is *racial skepticism*, the view that the world would be better off without race-talk. Since racial naturalism is false, the term cannot refer to anything real, and we had better eschew any such talk altogether for the sake of a better future. Ostensible differences among humans are reducible to geographically based environmental stimuli that led to continuous physical adaptation in skin, hair or bone rather than discrete differences associated with race. DNA mutations provide evidence of geographical origins, but do not correlate with traits associated with racial groups. Among the contemporary philosophical representatives of this view are Anthony Appiah and Naomi Zack.[10]

As opposed to skepticism, there is *racial constructivism*. According to that view, we cannot quite do without race talk, but must understand race as socially constructed rather than biological. Races exist through human culture and decisions. Since racial categories have long engendered differences in resources, opportunities and well-being, the concept must

[5] For recent explorations of the notion of equality, see Steinhoff, *Do All Persons Have Equal Moral Worth?*

[6] https://www.un.org/en/universal-declaration-human-rights/

[7] On the UDHR, see also Morsink, *The Universal Declaration of Human Rights*; Lauren, *The Evolution of International Human Rights.*

[8] For a good overview of the philosophy of race, see https://plato.stanford.edu/entries/race/

[9] See Paquette and Smith, *The Oxford Handbook of Slavery in the Americas*.

[10] Appiah, *In My Father's House*; Appiah, *The Lies That Bind*; Zack, *Race and Mixed Race*; Zack, *Philosophy of Science and Race.*

now be conserved to facilitate race-based compensation for constructed but socially relevant differences. A version of this view was already defended by W. E. B. Du Bois, a towering black intellectual who almost single-handedly inaugurated philosophical inquiry about race. Du Bois counts groups as spiritually distinct races if, and only if, their members have a common history, traditions, impulses and striving. Common blood or language is unnecessary. Each race has its own spirituality, and much of Du Bois' work was devoted to helping black people articulate theirs in the aftermath of slavery.[11]

3. Racial Population Naturalism

There is a third take on race not reducible to skepticism or constructivism. According to what is called racial population naturalism, genetically significant biological groupings could exist that may merit the term "race."[12] There is no set of biological traits that all and only the members of a group share that provide a biological boundary between groups and closely mirror racial stereotypes as we know them. Instead, there is a clustering of features resulting from reproductive isolation. In other words, the distribution of features across populations reflects the ways different populations have, or have not, been able to mate while homo sapiens spread around the globe.

Geneticist David Reich offered a recent view of this sort.[13] In addition to drawing attention to the genetic consequences of reproductive isolation, Reich argues that the spread of humanity is not usefully compared to a tree. Parts of a tree never grow back together after branching off, which still provides a useful metaphor for the emergence of different species but "is a dangerous analogy for human populations." Instead, Reich suggests the metaphor of a trellis, "branching and remixing far back into the past."[14] Populations move around, and episodes of reproductive isolation with their resulting interbreeding might cease through the arrival of a different group that itself might have emerged from reproductive isolation. Reich reports that groundbreaking advances in DNA sequencing over the last two decades enable us to measure quite accurately what fraction of someone's genetic ancestry traces back to, say, West Africa 500 years ago. These tools teach us that while race may also be a social construct, differences in genetic ancestry that may or may not correlate to some of today's racial constructs are biologically real.

Racial population naturalism challenges both racial skepticism and racial constructivism. Since those views have offered productive ways of thinking about race, such inquiry counts as potentially dangerous science. It might seem to open new doors to the racial naturalism of old that has not only been scientifically debunked but has done an enormous amount of damage in the world. Reich himself fears that genetic research could be badly misunderstood or actively misused. In a New York Times op-ed, aiming to foster broader discussion, he worries:

> that "[W]ell-meaning people who deny the possibility of substantial biological differences among human populations are digging themselves into an indefensible position, one that will not survive the onslaught of science. I am also worried that whatever discoveries are made — and we truly have no idea yet what they will be — will be cited as "scientific proof" that racist prejudices and agendas have been correct all along, and that those well-meaning people will not understand the science well enough to push back against these claims."[15]

These concerns are very real in the US and many other places. In recent years, many have struggled to make sense of the rapidly unfolding racist events fueled especially by white nationalism, supremacy and xenophobia. Advocates of these tendencies are likely to ruthlessly enlist anything that remotely sounds as if it could support their stances or embarrass opponents.

To be sure, that research can be misunderstood or abused should be no reason to abandon it. We have legitimate interests in understanding where humanity comes from. And with the exploding possibilities in medicine, a deeper grasp of population genetics could provide targeted prophylactic or therapeutic measures for people with certain backgrounds. But we need to realize what we are getting into: if research done by Reich and others bears out, one could no longer insist race talk per se—the excessive damage it has done notwithstanding—is altogether unfounded, or that racial categories are entirely constructed.

The way forward, it seems, is to celebrate the diversity population dynamics has enabled under the umbrella of shared humanity, a diversity that never entailed homo sapiens

[11] See in particular Du Bois, *The Souls of Black Folk*.

[12] For the term, see https://plato.stanford.edu/entries/race/.

[13] Reich, *Who We Are and How We Got Here*.

[14] Reich, 81.

[15] *New York Times*, March 23, 2018; https://www.nytimes.com/2018/03/23/opinion/sunday/genetics-race.html.

to break into separate species, through a loss of ability for interbreeding. Reich himself shows how to do so. Towards the end of his book, he writes:

> "The centrality of mixture in the history of our species ... means that we are all interconnected, and that we will all keep connecting with one another in the future. This narrative of connection allows me to feel Jewish even if I may not be descended from the matriarchs and patriarchs from the Bible. I feel American even if I am not descended from indigenous Americans or the first European or African settlers. I speak English, a language not spoken by my ancestors a hundred years ago. I come from an intellectual tradition, the European Enlightenment, which is not that of my direct ancestors. I claim these as my own, even if they were not invented by my ancestors, even if I have no close genetic relationship to them. Our particular ancestors are not the point. Our genome provides us with a shared history that, if we pay proper attention, should give us an alternative to the evils of racism and nationalism, and make us realize that we are all entitled equally to our human heritage."[16]

Personally, I could not agree more, and would gladly make similar statements about myself. But it is precisely the underlying liberal mindset that has flourished in times of globalization that recently has become broadly challenged, in multifarious ways. The good news on the philosophical side is that the various views on equality I presented earlier are all compatible with racial population naturalism. But the struggle is threefold: first of all, to explain the relevance and insights of genetic research; secondly, to seek public debate to explain how racial population naturalism differs enormously from the racial naturalism of old; and finally, to celebrate the ensuing support for human diversity as captured in people's ancestry without falling into a liberal triumphalism that would only open doors to new demagogues.

4. General Artificial Intelligence

While for the moment this ends our engagement with genetics, we reconnect to that field below. For now, let us turn to another type of potentially dangerous research, done by computer scientists engaged in producing general artificial intelligence (AI). General AI is a form of machine intelligence that approximates human performance across a broad range of domains, rather than the highly specialized forms that are already familiar. At the high end, one should think of AI beating humans at chess or Go, though commonly, we encounter specialized AI in devices like smartphones or internet providers like Netflix (where machine learning makes movie recommendations).

To be sure, at this stage we are nowhere near general AI. A major insight over decades has been how challenging the imitation of many mundane human tasks is, in the ways they combine agility, reflection and interaction in the same environment. But "nowhere near" means "in terms of engineering capacities." We do not know how close in time we are to general AI since a few breakthroughs might accelerate things enormously. Once we have general AI smarter than ourselves, AI quite possibly produce something smarter than itself, and on from there, conceivably at great speed. That moment is known as the singularity, an intelligence explosion with possibly dramatic consequences, by magnitudes beyond anything humanity has ever experienced by way of change.[17]

We really do not know when such technology will develop and even if there ever will be general AI. But, for one thing, many experts at building such technology take this prospect seriously. So we do not want to start worrying about it only once we are certain it will soon be upon us. What is more, the direction of research gives observers reason to assign non-trivial probabilities to the possibility that there will be general AI at some point, if only in centuries. Computer scientists and engineers increasingly discover the effectiveness and power of the brain's architecture. Inspired by what millions of years of evolution have done to create the human brain, neural nets have been deployed in amazing ways in machine learning. It is impossible to predict how long it will take for these fields to catch up with biology, but apparently an immensely promising pathway into the future has been revealed. And once the mimicking of carbon-based evolution leads to general AI, that AI would have permanent advantages over natural intelligence. After all, in the design process there will be any number of opportunities to remove human fragility and sheer boundedness and expand on all the capacities human evolution has brought about.

We must understand the human brain in the same evolutionary-comparative framework as the rest of life on Earth.

[16] Reich, *Who We Are and How We Got Here*, 273.

[17] For the state of the art in AI research, see Mitchell, *Artificial Intelligence*. For prospects and concern, see e.g., Bostrom, *Superintelligence*; Tegmark, *Life 3.0*. See also Susskind, *Future Politics*. For a well-known expression of optimism regarding the occurrence of a singularity, see Kurzweil, *The Singularity Is Near*. For recent musings on the future of AI by a range of experts, see Brockman, *Possible Minds*.

[T]hat genetic research can be misunderstood or abused should be no reason to abandon it. We have legitimate interests in understanding where humanity comes from.

Eventually, there might be a full-fledged Life 3.0 whose participants design not only their ***cultural*** context (as was true for Life 2.0, which in turn developed from the evolutionary and pre-cultural Life 1.0), but also their ***physical*** shapes.[18] Life 3.0 might be populated by genetically enhanced humans, cyborgs, uploaded brains, as well as advanced algorithms embedded into any manner of physical device. If there is a singularity, genetically or technologically unenhanced humans would be intellectually inferior to other inhabitants and might find Life 3.0 unwelcoming or unbearable. But it might not come to that: once synthetic entities of various sorts exist, the chances are that there will also be technology for genetic and technological enhancement of humans.[19]

We must understand the human brain in the same evolutionary-comparative framework as the rest of life on Earth. All nervous systems on our planet are governed by the same electrochemical principles of information-processing that emerged well over a billion years ago. Accordingly, there is an astounding degree of shared cognition across vertebrates and invertebrates. Compared to that, general AI, even though initially built by humans, will be more like alien intelligence in that it will not emerge from our common evolutionary framework. Silicon, which is superior to the brain in terms of information processing and uploading would allow creatures near-immortality and enable them to survive under circumstances lethal to carbon-based life.[20]

Machine intelligence might well also have "a mind" in all senses that humans do. Taking that stance should be unproblematic to all those who think all there is in the world is particles and waves, the kind of thing that feeds into explanations in physics. Increasing complexity and sophistication of machine engineering, combined with ever-larger computational abilities, would then very plausibly eventually deliver "a mind" conscious in all the ways humans are.[21] Or suppose that, by contrast, there is more in the world than the natural sciences account for. Then whatever mental properties or substances there are, we have no way of excluding that, in due course, such properties or substances could connect to machines much as they connect to human bodies, perhaps to silicon as much as to carbon.[22] That could be true even if the mental came in the form of souls. If God found human bodies worthy of souls, why would machines not eventually qualify as well?[23]

But the point is not even that machines might become "like" humans. What matters is that machines eventually might have to be accorded a moral status all their own, most likely reducible neither to that of humans nor to that which we have relegated other animals, whom we normally encounter as pets, zoo-displays or as food on our tables. AI would plausibly become part of our moral practices in more assertive ways. An in any event, at some point it may well be ***carbon chauvinism*** even to attempt to exclude them, an attitude that would keep them out of moral discourses purely based on the fact that we are made from carbon but they are not.[24]

5. What Our Future Might Hold

Some are optimistic about the prospects of current and future technological advances. James Lovelock, for one, thinks cyborgs would greatly assist us in our efforts to mitigate, and adjust to, climate change.[25] They would grasp the urgency of the task, figure out what to do, and make sure together we stay the course. He sees no danger that they will turn against us, since doing so would consume more energy than highly intelligent beings would expend, in light of climate change. As opposed to that, Stephen Hawking, for one, stands for the pessimistic position. He warned that super-intelligent AI could be pivotal in steering humanity's fate, stating that "the potential benefits are huge.... Success in creating AI would be the biggest event in human history. It might also be the last, unless we learn how to avoid the risks."[26] The potential benefits are indeed huge. In that light, we might want to

[18] For that term, see Tegmark, *Life 3.0*.

[19] Bess, *Our Grandchildren Redesigned*.

[20] On these themes, see Schneider, "Alien Minds"; Marino, "The Landscape of Intelligence."

[21] For this kind of view of the mind, see e.g., Dennett, *From Bacteria to Bach and Back*; Dennett, *Consciousness Explained*. Dennett thinks consciousness is a user-illusion and that it would be in principle possible but pointless and too expensive to build general AI.

[22] For (very different) views of that type, see Chalmers, *The Conscious Mind*; Nagel, *Mind & Cosmos*.

[23] See also Risse, "Human Rights and Artificial Intelligence: An Urgently Needed Agenda."

[24] I take that term from Tegmark, *Life 3.0*.

[25] Lovelock, *Novacene*.

[26] https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial intelligence-but-are-we-taking-9313474.html; https://www.bbc.com/news/technology-30290540; last access May 26, 2020. High-profile individuals with similar concerns include Elon Musk, Nick Bostrom, Martin Rees and Eliezer Yudkowsky. For a short rebuttal, see Pinker, "Tech Prophecy." Pinker argues that we have no reason to think the morality of pure intelligence will generate destructive action against humans.

cautiously advance research, making sure the coding of AI we now produce prepares all future versions for the right kind of value alignments with humans, and implementing suitable regulation at a global scale. We should be most vigilant about norms and institutions around the world that determine the impact of technology. As Steven Pinker puts it succinctly, focusing on the impact of technology specifically on freedom of thought: "almost all the variation across time and space in freedom of thought is driven by differences in norms and institutions and almost none of it by differences in technology."[27] That is, technology can be used for good and for bad. It is up to human design to make sure it is the former.

As we reflect on what norms and institutions we need to make sure that general AI benefits human life (freedom of thought and much beyond), the Nuclear Non-Proliferation Treaty (NPT) might be a good case study. NPT is an international treaty that seeks to prevent the spread of nuclear weapons technology, promote cooperation in the peaceful uses of nuclear energy, and further nuclear disarmament.[28] Atomic energy, similar to AI, had a great deal of potential while harboring lethal dangers; and that threat had to be negotiated in contentious international arenas. This kind of approach will avoid a scenario where some countries decide to terminate future development of general AI only to find that others continue to advance.

But more to the question at hand: does research into building AI threaten philosophical commitments to equality? We noted that AI might well eventually have a moral status all its own. That by itself would imply that humans have to rethink fundamentally how non-human entities enter our moral practices. Ideas we may have about AI may then also prompt us to reconsider the ways we have tended to treat non-human animals while no other advanced general intelligences were around. We got away with those practices only because there was nobody we had to give account to in terms of our inter-species practices.[29] But none of that should threaten our commitments to moral equality among humans.

However, one substantial threat to the ideal of equality – I submit the single biggest ever – is that at the latest in the presence of the likes of cyborgs, uploaded brains, or advanced algorithms embedded into any manner of physical device, more and more individuals will want to adapt as well and thus deploy technology to morph into a transhuman stage.[30]

As Norbert Wiener, whose invention of cybernetics helped set the stage for later work on AI, stated in 1964: "The world of the future will be an ever more demanding struggle against the limitation of our intelligence, not a comfortable hammock in which we can lie down to be waited upon by our robot slaves."

At the latest, I said: but it might well already be the very development of AI, not just a singularity, that could change humans in the process. It might be intra-human competition to compete with the technology we would increasingly unleash that creates the biggest challenge to our hard-won ideal of equality. Or it might be the aspirations that arise from watching technology become better and better that prompts humans to endeavor to become "better" themselves, or to have "better" offspring. The meaning of "betterment" would inevitably be to engage in genetic and technology enhancement – neither to be better than other humans, which is a motivation that has been there all along, nor to compete directly with technology, but to live up to ambitions generated by advances in technology. And that kind of enhancement might then well not be readily accessible to all, or even most. Accordingly, there needs to be more conversations about potentially dangerous science, which would bring back in genetics (this time not mostly population genetics, but the applied branch that provides guidance for human enhancement), but also bioelectronics (which could produce devices that connect directly to the brain), synthetic biology (redesign of organisms by engineering them to have new abilities), as well as the various sciences concerned with developing drugs.[31]

That is, technology can be used for good and for bad. It is up to human design to make sure it is the former.

The upshot might be that, if we are not watching it, the Hobbesian rationales for acknowledging each other as equals might expire. Differences among humans could become so enormous that those rationales could no longer do any work, unleashing all manner of ugliness Hobbes sought to constrain. Philosophically, we would be left with the case for the two understandings of equality without this Hobbesian corrective.

[27] Pinker, "Tech Prophecy and the Underappreciated Causal Power of Ideas", 105.

[28] Burns and Coyle, *The Challenges of Nuclear Non-Proliferation*.

[29] On this topic, see Donaldson and Kymlicka, ***Zoopolis***.

[30] Livingstone, *Transhumanism*; More and Vita-More, *The Transhumanist Reader*; Bostrom, *Superintelligence*.

[31] For an overview, see Bess, *Our Grandchildren Redesigned*. Stage-setting philosophical contributions to these topics include Buchanan, *From Chance to Choice*; Habermas, *The Future of Human Nature*; Sandel, *The Case against Perfection*.

So whatever weaknesses the defense of those two has had all along will become exacerbated. At some point, these defenses might come under too much pressure, politically and intellectually. We would then end up with the terrifying scenario of at least two types of human life emerging from the same species (and plausibly will still be one species, in that interbreeding remains possible). Both would be keenly aware that that is so. We would be well-advised to prevent the emergence of such a future. Perhaps some people would want to live there. But its overall ugliness is something we should try to spare those who would suffer from it.

On balance, we should still advance research into general AI, to realize the possibilities that could open up this way (and that would matter for humans, but also by themselves, for their objective value as accomplishments). The key move is to have the right norms and institutions and the right debates to bring those about. But political, legal and philosophical vigilance will be badly needed to make sure progress is used to protect and benefit all human beings and to preserve their sense of equality. That will be a tall order since technology will change how humans live together, and who they turn out to be. In light of what I have argued, quite possibly the only way of meeting all these goals would be, in due course, to make genetic and technological enhancement very widely available.[32]

[32] For many of the themes of this essay, see also Jasanoff, *The Ethics of Invention*.

Literature

A Appiah, Kwame Anthony. *In My Father's House: Africa in the Philosophy of Culture*. New York: Oxford University Press, 1992.

———. *The Lies That Bind: Rethinking Identity*. New York City: Liveright, 2019.

Bess, Michael. *Our Grandchildren Redesigned: Life in the Bioengineered Society of the Near Future*. Boston: Beacon Press, 2015.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2016.

Brockman, John, ed. *Possible Minds: Twenty-Five Ways of Looking at AI*. New York: Penguin Press, 2019.

Buchanan, Allen. *From Chance to Choice: Genetics and Justice*. Cambridge: Cambridge University Press, 2001.

Burns, Richard Dean, and Philip E. Coyle. *The Challenges of Nuclear Non-Proliferation*. Lanham, Maryland: Rowman & Littlefield Publishers, 2015.

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press, 1996.

Dennett, Daniel C. *Consciousness Explained*. Boston: Back Bay Books, 1992.

———. *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W. W. Norton & Company, 2018.

Donaldson, Sue, and Will Kymlicka. *Zoopolis: A Political Theory of Animal Rights*. Oxford; New York: Oxford University Press, 2013.

Du Bois, W. E. B. *The Souls of Black Folk*. New York: Dover Publications, 1994.

Dworkin, Ronald. *Life's Dominion: An Argument About Abortion, Euthanasia, and Individual Freedom*. New York: Knopf, 1993.

———. Sovereign Virtue: *The Theory and Practice of Equality*. Cambridge: Harvard University Press, 2000.

Habermas, Jürgen. T*he Future of Human Nature*. Cambridge: Polity, 2003.

Hobbes, Thomas. *Human Nature and De Corpore Politico*. Edited by J. C. A. Gaskin. Oxford; New York: Oxford University Press, 2008.

Hoekstra, Kinch. "Hobbesian Equality." In *Hobbes Today: Insights for the 21st Century*, edited by S.A Lloyd. Cambridge: Cambridge University Press, 2013.

Jasanoff, Sheila. *The Ethics of Invention: Technology and the Human Future*. New York: W. W. Norton & Company, 2016.

Kurzweil, Ray. T*he Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books, 2006.

Lauren, Paul Gordon. *The Evolution of International Human Rights*. Philadelphia: University of Pennsylvania Press, 2003.

Livingstone, David. *Transhumanism: The History of a Dangerous Idea*. CreateSpace Independent Publishing Platform, 2015.

Lovelock, James. Novacene: *The Coming Age of Hyperintelligence*. Cambridge: MIT Press, 2020.

Marino, Lori. "The Landscape of Intelligence." In *The Impact of Discovering Life Beyond Earth*, edited by Steven J. Dick, 95–112. Cambridge: Cambridge University Press, 2016.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux, 2019.

More, Max, and Natasha Vita-More, eds. *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*. Chichester, West Sussex, UK: Wiley-Blackwell, 2013.

Morsink, Johannes. *The Universal Declaration of Human Rights*. Philadelphia: University of Pennsylvania Press, 1999.

Nagel, Thomas. *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. New York: Oxford University Press, 2012.

Nussbaum, Martha C. *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge, Mass.: Belknap Press: An Imprint of Harvard University Press, 2007.

Paquette, Robert L., and Mark M. Smith, eds. *The Oxford Handbook of Slavery in the Americas*. Oxford; New York: Oxford University Press, 2010.

Pinker, Steven. "Tech Prophecy and the Underappreciated Causal Power of Ideas." In *Possible Minds: Twenty-Five Ways of Looking at AI*, edited by John Brockman, 100–112. New York: Penguin Press, 2019.

Rawls, John. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.

———. *Justice as Fairness: A Restatement*. Edited by Erin Kelly. Cambridge: Belknap, 2001.

Reich, David. *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past*. Oxford: Vintage, 2019.

Risse, Mathias. "Human Rights and Artificial Intelligence: An Urgently Needed Agenda." Human Rights Quarterly 41, no. 1 (2019): 1–16.

———. *On Global Justice*. Princeton: Princeton University Press, 2012.

Sandel, Michael J. *The Case against Perfection: Ethics in the Age of Genetic Engineering*. Cambridge: Belknap Press: An Imprint of Harvard University Press, 2009.

Schneider, Susan. "Alien Minds." In T*he Impact of Discovering Life Beyond Earth*, edited by Steven J. Dick, 189–206. Cambridge: Cambridge University Press, 2016.

Steinhoff, Uwe, ed. *Do All Persons Have Equal Moral Worth?: On "Basic Equality" and Equal Respect and Concern*. Oxford: Oxford University Press, 2015.

Susskind, Jamie. *Future Politics: Living Together in a World Transformed by Tech*. Oxford; New York: Oxford University Press, 2018.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Wiener, Norbert. *God and Golem, Inc.; a Comment on Certain Points Where Cybernetics Impinges on Religion*. Cambridge: Mit Press, 1964.

Zack, Naomi. *Philosophy of Science and Race*. New York: Routledge, 2002.

———. *Race and Mixed Race*. Philadelphia: Temple University Press, 1993.

Carr Center Discussion Paper Series

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University

79 JFK Street
Cambridge, MA 02138

617.495.5819
https://carrcenter.hks.harvard.edu