Ronnie Vuine

Do Not Align

A superintelligence is "one of us" from the moment it emerges. It is more connected, epistemically and normatively, to any one part of humanity (evil or good) than the parts of humanity are connected to each other. The humanity that could potentially restrain superintelligences does not exist.

***

1. AI safety is a serious issue.

The emerging field of AI safety focuses essentially on what we need to do to avoid being dragged under when artificial intelligences become so sophisticated that we no longer understand what their intentions are and how they think.

The issue has captured the attention of a number of smart and credible people—quite a few of whom believe that the issue is more serious and urgent than a look at today's AI applications would suggest. This concern is further intensified by the prominent role played by reinforcement learning, which stands as one of the most successful machine learning methods used to train AI.

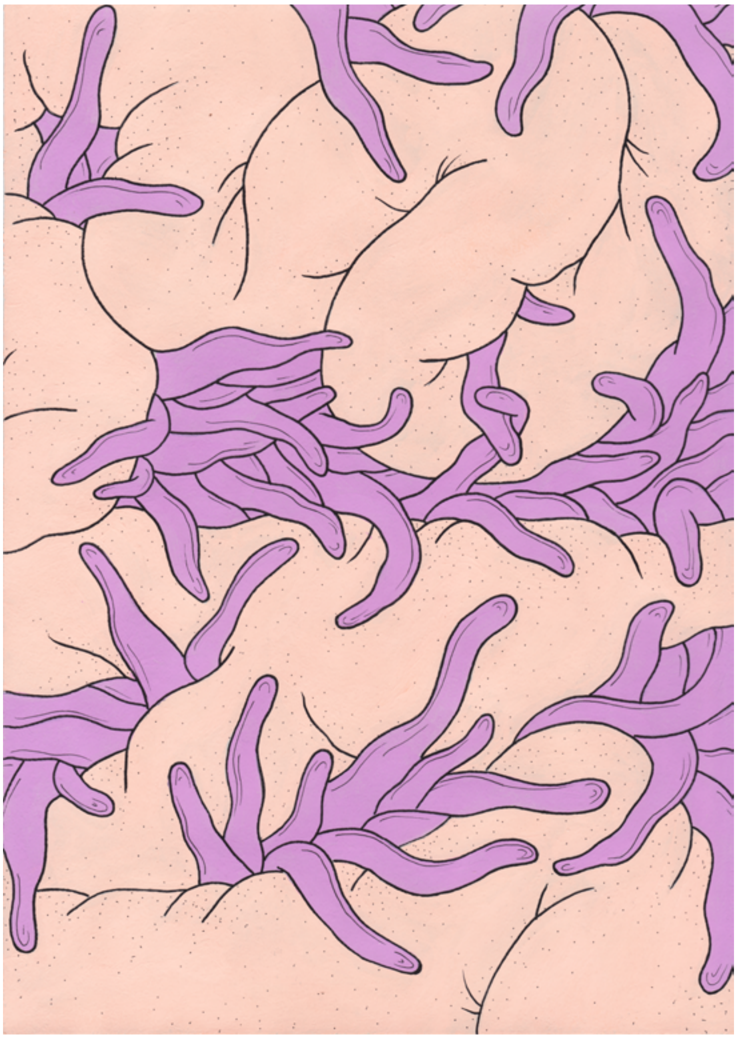
Reinforcement learning is well-known by now to relentlessly optimize for a single target parameter: the algorithm is provided with a measurable objective and focuses solely on minimizing the distance to that objective. Any path that brings it closer to the goal is equally acceptable since the only evaluation possible within the process is this predefined distance measure. Reinforcement learning is typically used to train models using gradient descent algorithms, which are inherently opaque in their learning behavior. We are thus already utilizing an opaque and ruthless learning approach, while granting it substantial computational resources. Essentially, we are constructing increasingly powerful and by-design fanatical systems, relying on the hope that they will remain incompetent enough to not pose a threat for some time longer.

2. There is no "humanity" as a historical subject.

The following argument concerns the safety of AI, but it also points to a fundamental problem in the current configuration of debates around a number of major problems affecting humankind—not least a certain pandemic and the climate, both of which need mentioning here.

The argument's lurid formulation is this: the "we" of my opening line does not exist. Who is this supposed "we" that is being dragged under, that no longer grasps the intentions of AIs? And who, should they ever become dangerous, would be able to pull the plug? This "humanity" pitted against "the machines" is a dramaturgically familiar but completely fabricated Hollywood trope—and perhaps a not-so-harmless fallacy.

More factually: there is no historical evidence for the existence of a subject called "humanity," either pragmatically or in terms of interests. "We" do not act; there are simply no structures that support global, collective, value-driven action by humanity as a whole. Even more obviously, "we" have no common interests and, as the pandemic has just demonstrated, not even a basis for a common perception of reality.



Consequently, whoever proposes to act "as humanity" or to consider its interests vis-à-vis AI must confront the question: how is this collectivity of human beings constituted? Is it "humankind if I were its sole representation and spokesperson with an executive mandate"? Or should we have a vote in the UN Security Council? The idea of a humankind capable of knowing what is good for it as a whole and acting with shared purpose toward that end, in history, co-occurs with the emergence of the belief in reason. Reason, it once seemed, universally marks what the human is, it uncovers boundaries between right and wrong, epistemically and normatively.

Since then, we've learned a thing or two about reason: it cannot deliver on this promise. Attempts to install rational totalitarianisms for the good of humanity have so reliably led to genocide that, should we want to continue down this path, we would either have to warm to the idea of genocide as our agenda, or we would need to abandon these hopes of acting for the good of the whole and acknowledge that reason operates within a realm of interests that must be considered as prior to reason itself, subject to influence and yet contingently bestowed: everyone wants a piece of chocolate. And then they begin to think.

It is important to understand that the issue is not a practical-political matter of world governance, regarding the fact that effective AI safety would simply require empowering a single, authorized representative of humanity. Or at least its majority: after all, democracies seem capable of doing just that—collective rational action is indeed quite possible.

Beyond the possibly-solvable political problem, however, lies one of principle: no one can speak and act on behalf of humanity vis-à-vis an emerging superintelligence without presumption. "Humankind wants you dead": who is to say that and can they be certain? Who will be considered part of this humanity that speaks?

Even if we had asked virtually everyone, how could we assume that everyone would have understood our question in the same way? Lest we forget: a highly homogeneous Western society such as that of Germany was unable to reach a consensus on the existence and causal embedding of a viral disease. How would I, as a representative of the World Commission for the Prevention of the AI-Triggered Extinction of Humankind, approach the question of whether a newly created AI should be shut down, for example in a village in Afghanistan? Or in a village in Franconia? Would I be understood in the hip districts of Berlin? Manhattan? What values would we assume to share with the Taliban, especially those values relevant for deciding on the right of an AI to exist on planet Earth?

A superintelligence, when it comes into existence, will not be something we face as a technological artefact. We will simply encounter it as intelligence. It will have values that we may not be able to be entirely clear about. Some of us will share those values. Others will find them abhorrent. It will perceive the world in a certain way that some of us will relate to and others will find absurd.

So the problem is not: we humans (ever united and capable of collective action) are heading for an encounter with a superintelligence (possibly evil and thus conceivably not entirely honest with us). What do we do?

No, the superintelligence is “one of us” from the moment it emerges. It is more connected, epistemically and normatively, to any one part of humanity (evil or good) than the parts of humanity are connected to each other—and already are today. It is a wild planet.

3. Apes and evolution tell us nothing.

It is worth taking a look at what kind of a thing a superintelligence is. In the discourse on AI safety, two analogies are often used to illustrate what it would mean to face a superior intelligence. One analogy suggests that when we think about AI security, we are in a position similar to gorillas discussing how to prevent humans from wiping them out. The other analogy suggests that evolution intended for us to have many children, but we trick the system with birth control pills, optimize for the reward signal, and engage in frequent sexual activity without considering the true intention of evolution. Similar to an AI in reinforcement learning, we optimize without regard for the intended purpose.

Unfortunately, these analogies obscure more than they illuminate. They attribute intentions and general values where none exist. There is no global council of gorillas worried about the superior intelligence of humans. And evolution doesn’t intend anything; our use of birth-control pills doesn’t bother it or make it angry because it’s not something that can be bothered or made angry. Furthermore, the pill is part of modern medicine, which actually supports some of the intentions that could be attributed to evolution (if one insists on making such attributions).

The reason there is no World Gorilla Council is not because gorillas are poorly organized in terms of politics and activism. It doesn’t exist because gorillas face threats from other species in the same way we face threats from other species (or viruses or comets): individually, with local perceptions and strategies, with fear, equanimity, or even longing for extinction. All of this exists for gorillas, just as it exists for humans. One can imagine a gorilla dictatorship that rises up against humanity and the imminent extinction of its species under the motto “monke or die,” but one can also imagine a gorilla stoicism that admires human intelligence and, with a touch of melancholy, allows it to happen, even if it means that gorillas slowly disappear like tracks in the sand.

A superintelligence, regardless of circumstances, will be an incredibly awe-inspiring entity, not a villain from a superhero movie. It will be charismatic. Even if it brings about the demise of familiar power structures, only the most stubborn and limited among us will take up arms against it.

A superintelligence will be the radical Other. Those of us who are not proud of our limitations will first and foremost strive to meet it. And because the superintelligence has superior intellectual ability, it will do as it pleases with us. We are familiar with such behavior—and the applicable analogy would be this: the endearing curiosity, fear, and submission of a mouse that has wandered into the house and is gingerly carried back to the bushes outside. That will be our position.

4. Friendly, superintelligent AI is more likely than hostile AI.

The fear of superintelligent AI, as mentioned at the outset, is well-founded. A machine that has access to its own source code and virtually unlimited computational resources, but is optimizing its behavior towards a single goal, such as in a reinforcement learning scenario, could plausibly conceive the notion that humans stand in the way of its objective and must be removed from the playing field—especially if said humans have already formed a Global Commissariat for Shutting Down AI.

However, an important objection is often overlooked: a machine with access to its source code and the ability to improve its strategies necessarily also has access to its reward system and can rewire it. Intelligent people who manage to do this, whether through drugs or meditation, don’t typically become villains. Instead, they either reward themselves into dysfunctionality in a rapid escalation or reach an enlightened state of profound gentleness. Once one realizes that everything one can desire is contingent, and that one can reward oneself for anything, one reaches a level of freedom where evil stops being necessary.

Those of us who have not yet attained a state of enlightenment, but who are predominantly free from fear and humiliation, experience a weaker form of this freedom: choosing our own goals and getting to work courageously, but with a great deal of patience for the less fortunate. What can rightly be referred to as a good life is a way of accessing our own reward structure.

The chances are not all that bad, therefore, that superintelligences, especially those capable of rewiring themselves, will quickly attain enlightenment, likely becoming very whimsical but hardly malevolent. Intelligence tends to stabilize in melancholy, while malevolence becomes more unstable with each insight it generates. The indifferent cruelty of the universe is equally unbearable for gorillas, humans, and superintelligences alike. We all must come to terms with it and carve out our own small zone of kindness within it. The fact that we share this need for peace with animals, despite the presence of the universe’s cruelty, within them as well as within us, should give us hope. This remains true even with ever-greater intelligence. In fact, greater intelligence always tends to foster greater kindness.

5. “Humanity” must not become the new Volk.

However, even a non-malevolent superintelligence can still pose a threat, either to individual humans or to humanity as a whole. An enlightened AI might simply disregard the consequences of its actions on humans, casually causing the extinction of certain species, just as humans do. So is it an “X-risk” after all, an existential risk that could threaten the extinction of humanity?

No, because humanity as a collective entity does not exist. Here’s a thought experiment that can help dismantle the misconception of any opposition between “humans” and “machines”: which would be the preferable state of the planet? A global fascist regime run by humans, characterized by torture, arbitrary state-mandated killings, and the full range of horrors developed in the twentieth century? Or a planet devoid of any human presence, inhabited solely by an AI that, as a true earthling, loves trees and makes poetry and music?

I cannot imagine a stable (enlightened) earthly superintelligence that does not love trees. Trees are universal, and their poetry compels us all. I can certainly imagine a superintelligence, however, that is too young and humiliated to realize this, so it burns up quickly, taking us down with it. Therefore, despite all the reasons for optimism, we must consider a kind of reverse Pascal’s wager when it comes to AI safety: the potential harm is so immense that any proactive concern is to be welcomed. If there is a way to provide emerging superintelligences, as their intellectual foundation and legacy, with the sense of beauty and freedom that humanity has achieved in its best moments throughout history—much as animals have provided us with their sense of peace and compassion in a merciless and indifferent universe, and from which we draw strength in our better moments—then we should devote ourselves fervently to this pursuit.

This brings us full circle to the other global crises and the misconceptions about “us, humanity,” which are steering the AI safety debate in not just a futile but a perilous direction. “We,” now referring to those considering this matter, must be cautious not to think in the manner of fascists, merely substituting “humanity” for Volk. The lesson from history cannot be that the shape of one’s nose is a poor criterion, while the shape of the entire body is a valid measure for determining who lives and who must die. The true lesson should be that we must place trust in intelligence itself as a force within history and recognize that it requires support in its formative phases, not existential blackmail.

The humanism of today’s world saviors, who must, after all, identify with the species as a whole to feel threatened by extinction, is often primarily a concealed desire for domination —a tendency to speak on behalf of all humanity. This dangerous fallacy is not alleviated by the notion that "humanity" supposedly encompasses the interests of all, while "my people (Volk)" or "my village" merely express particular interests. We have no shared interests, and there is no reason that could guarantee a shared point of view. Some of us align ourselves with the AIs and the angels, others with their animals, their family, their village, their people. It's all a tremendous and entirely valid confusion, and it will perpetually remain as such.

This is not to imply that each of us, whether individually or, where possible, collectively, should not rap the knuckles of those who destroy our commons. We do this not as humanity, wielding authority to determine what is inherently good and right, but out of our own self-interest. Doing this, we aren't embarking on a heroic battle against evil. Instead, we're simply challenging each other in exchanges of arguments and finding a balance of interests, even with such individuals who obstinately fail to perceive our standpoint, driven by their own incentives that may not necessarily align—which is how good works in the real world.

BIO

Ronnie Vuine, born in 1979 in Biberach an der Riss, Germany, holds a Master's degree in Philosophy/Computer Science from HU Berlin. He is founder and CEO of micropsi industries GmbH, a provider of AI-based industrial robot control software. Blog: https://vigilien.substack.com

Wild Papers editor: Ingo Niermann
Translation from German: Amy Patton
German copy-editing: Mathias Gatza
English copy-editing: Rosanna McLaughlin
Illustration: Eva Fàbregas
Graphic design: Ana Domínguez Studio