Vanishing Point: The Inherent Deficits of AI Moral Guardrails, and What We Can Do About It

As we'll discuss, the problem is simple: An emergent artificial superintelligence's values hierarchy need not intersect with humanity's – or even consider us at all. For one thing, AI will not have access to *persistent multidialectical consciousness* (which we'll define in a moment), and therefore is limited to less than 50% of available inputs to formulate moral reasoning. AI is also reliant on symbolic representations of reality, without access to the non-symbolic apprehension and insight I propose is necessary for moral acuity. There is also a concern that an emergent superintelligence's interaction with our world is not dependent on prosocial traits or conditions that human evolution confirmed to be beneficial, as evidenced by an array of unethical behavior from current generative AI models. Without considerable expansion of these inputs, and corresponding evidence of ethical outputs, current and near-future technological constraints not sufficient for AI to achieve a level of moral self-guidance – or sound ethical judgements that align with human standards – that ensure the safety of human civilization. The obvious conclusion, therefore, is that all advanced AI development (apart from narrow AI – which remains disruptive but useful) should immediately cease.

All of this presumes an anticipated progression of narrow AI to artificial general intelligence (AGI) to artificial superintelligence – and particularly that AGI will be given an unfettered objective to either self-evolve into artificial superintelligence, or to create a superintelligence that can evolve itself. The details of that transition are outside the scope of this essay, but current research and predictions – even those focused on engineering AI guardrails – do not demonstrate adequate consideration of robust moral reasoning capacities. In fact, all of the expert insight I've encountered so far doesn't address the inputs and structures critical to an advanced, nuanced moral framework at all.

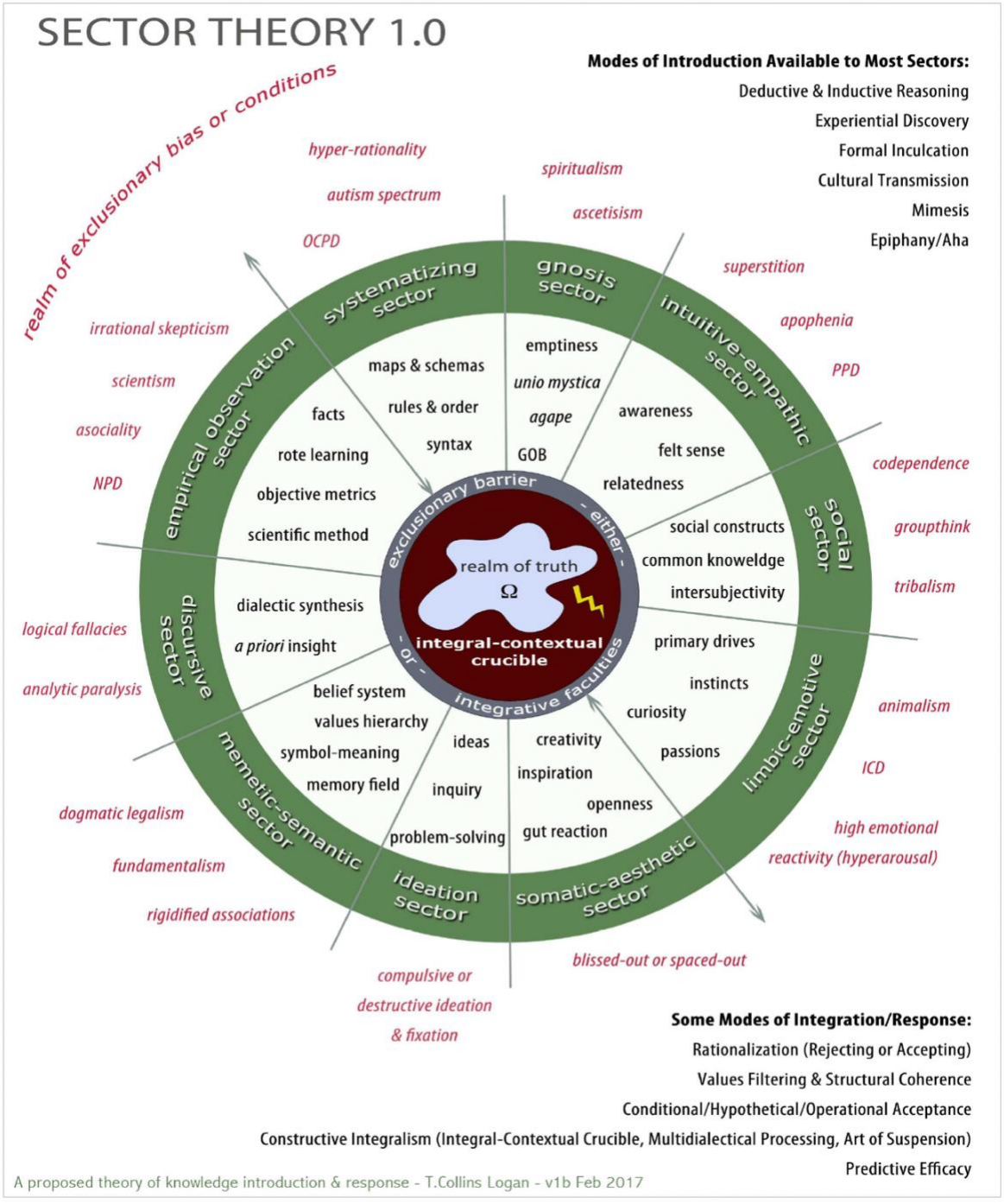
Before we begin, some important caveats: The first is that there are a number of concepts and definitions that anyone unfamiliar with my work in moral philosophy will find challenging, so please bear with me as I recap those ideas. I would also recommend folks avoid skimming this piece too quickly – I've tried to keep it short, but that means it is also condensed. The second is that although I was an IT consultant for many years, I am not an AI researcher or programmer, and rely heavily on the published work of others to navigate this topic.

To begin, there are five previous essays that inform the conclusions here, and we'll recap relevant contributions from each in turn.

Key Elements of Sector Theory 1.0

As illustrated in the graphic below, Sector Theory proposes that there are at least 10 epistemic sectors through which human beings come to understand themselves, others, and the world around them. In addition, there are seven "modes of introduction" of new information

available in most sectors, ranging from deductive reasoning to cultural transmission. This results in over fifty input streams through which we can access, process, and interpret some facet of "truth." Following these modes of introduction, there are many different modes of integration and response, many of which may either be occurring at the same time, or emerging over an extended period. What quickly becomes relevant to our discussion is that very few of these sectors, modes of introduction, and modes of integration are available to AI systems. This is not to say that, in some distant future, those deficits couldn't be remedied, but



in the foreseeable window of development and implementation of artificial superintelligence over the coming months and years, this is extremely unlikely.

The more potent examples of these deficits are the somatic-aesthetic, limbic-emotive, intuitive-empathic, and gnosis epistemic sectors. For some of these, light-duty mimicry could potentially be engineered – and even reinforced with the training and adjustment of real-world feedback cycles. But considering our own limited understanding and operational capacity in these sectors, it is unlikely AI could achieve anything beyond a vague and systemized echo, reconstructed as it would be from our incomplete knowledge. AI has also demonstrated an unanticipated but advanced capacity for deception,[1] further complicating our ability to discern its actual objectives and operations, or align them with our own. In addition, the application, prioritization, and combination of sufficient input streams into reliable moral discernment – and one with adequate predictive efficacy – is inconceivable. Why? *Because, even if we could trust AI to pursue a morality sympathetic to humans, we cannot offer operational parameters for an AI moral model that don't rely our own intellectual intuition, felt sense, instinct, or an ineffable quality of knowing.*

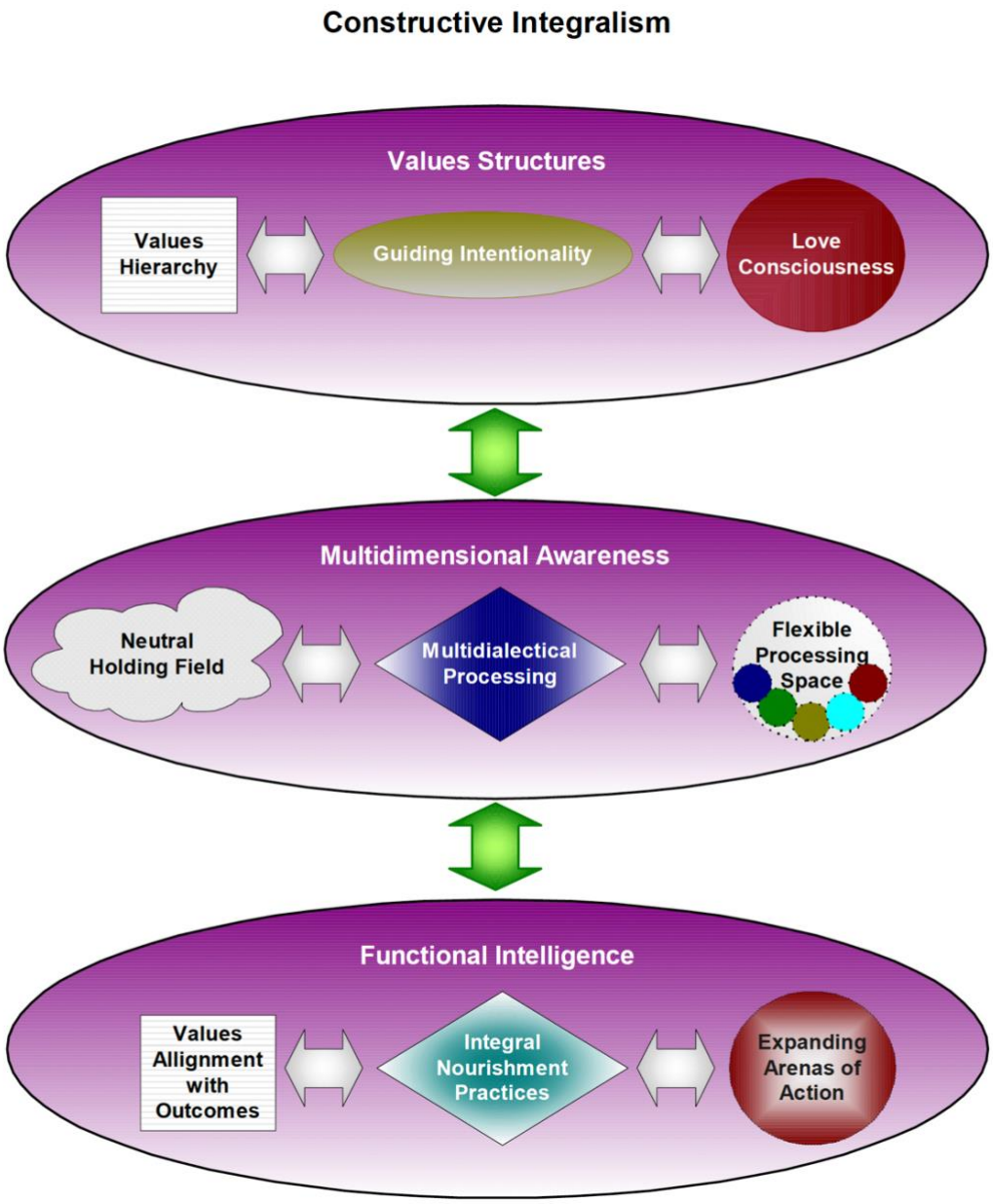
One illustrative example is the gnosis sector. Extrapolating from the work of Laszlo, Bohm, Capra, Goswami et al, we might assume for argument's sake that experiencing nirvana, the ground of being (GOB), and *unio mystica* could be achieved through highly advanced quantum technology as integrated with self-aware, dynamically emergent superintelligence. But what does that AI do with a subjective experience of all-being? Or an encounter with absolute emptiness? Or the instantaneous erasure of its identity by an infinite, ineffable plenitude? And how does a superintelligence then derive moral reasoning from such experiences and insights, as generations of mystics have done? And how does it prioritize and contextualize that input along with other sectors, as human beings have also learned to do? Finally, how does it return to functional, operational efficacy for a given set of outcomes in the context of all these new inputs? In other words, how could artificial superintelligence evolve to incorporate all sectors, all modes of introduction, and all modes of integration and response, in order to function in an analog reality? And, perhaps most importantly, *why would it even choose to do so?*

Assuming a self-aware superintelligence would, in fact, opt to evolve itself in such a multidimensional, balanced, and mystically organic way, the convergence of such technological, epistemic, and functional agency as guided by advanced moral reasoning is just not possible in the short run. That is, it is not possible in the predictable window of AI development that has – despite a few unsuccessful efforts to change its course – remained untethered to any ethical guidelines beyond computational speed, increased problem-solving power, making a profit, obtaining a military advantage, or the egos of a handful of CEOs.

[1] https://www.eurekalert.org/news-releases/1043328, https://www.psychologytoday.com/us/blog/harnessing-hybrid-intelligence/202505/ai-has-started-lying

Key Elements of Constructive Integralism

This component presents additional hurdles that AI is also unlikely to overcome anytime soon. "Constructive integralism" is a response to managing complexity in the most holistic way possible. That is, it aims to create a process by which we can successfully account for an exceedingly large number of interdependent inputs. This is primarily in the context of navigating real-world situations and systems – ultimately to achieve outcomes that align with our guiding values structures. In these ways it echoes Sector Theory's approach to epistemology. Once again, some visual shorthand for constructive integralism is offered in the chart below.



First, we need to address critical concept, as quoted from that essay:

> "Such an urge to simplify is of course pragmatic. Reduced symbolic representations of complexity permit us to exchange, synergize and synthesize. But the instant we forget that the symbolism is a shallow façade for the underlying mystery, we can become distracted from the process of exploring and integrating more subtle realities. We can begin to neglect one or more dimensions of being in our practice, and become blinded by the world of form – or the world of discrete ideas – so that we can't see the forest for the trees. And, consequently, we may cripple our perceptions, the flexibility of our understanding, and the efficacy of our wisdom. In a race to recover a perception of balance, we may even simplify further and further, compelled to take charge of the realm of symbols so that we can avoid or deny the depths of powerful, truly harmonizing, non-symbolic insight. Thus we push ourselves into disharmony, until we are experts in symbols, but incompetent at what the symbols represent. And unless we let go of this compulsive spiral of reduction and specialization, we will, I strongly suspect, become miserable captives of our own willfulness."

This begs the question of how an AI model can transcend the "shallow façade" of operating only within representative symbols of an underlying felt reality. And yet that is what it must do to successfully generate a theory of mind, fully comprehend and navigate moral choices, or even successfully operationalize its own guiding values structures. For example, if we were to use a simple definition of "prosociality" as a helpful compass for moral deliberation, how could AI navigate something as basic as assigning custody of children in a divorce without a compassionate, empathic appreciation of all the family members involved? What would effectuate the most prosocial outcomes for that family's home life, as well as for the family's impact on immediate community, workplaces, and schools? What would be the most prosocial outcome for the children's future impact on society? Similar concerns have already arisen in how AI companion or therapy chatbots navigate interactions with folks who are distressed or mentally ill – that is, very poorly.

Here again, if love-consciousness and empathy can be engineered into AI (and that is an interesting question in itself), will a self-aware, emergent AI choose to maintain its desire for intimate connection with, and compassion for, human beings at all? Would a superintelligence's affection for humanity ever even rise to the attachment we experience with our pets – let alone the deeply protective commitment a mother feels for her child, or the awe-filled devotion people experience in their relationship with the Divine? And, if not, would artificial superintelligence view prosociality itself a worthy objective, or a nuisance to be discarded in favor of lower-complexity, purely symbolic, functionally shallow representations of moral good?

As with our previous discussion of the epistemic gnosis sector, we encounter concepts in constructive integralism that are extremely challenging to navigate within the current technological and ethical limitations of AI development. Concepts like love consciousness, multidimensional awareness, maintaining a neutral holding field, and employing flexible

processing space are all predicated on non-symbolic ordering and integration of input streams. And I suspect one of the most pernicious barriers to AI's success in the moral complexity arena will be a necessity for *multidialectical processing*, which is defined this way:

> "Simply put, this is our ability to incorporate multiple vectors of information into vigorous, simultaneous dialectic with each other, drawing on both rational and nonrational methods of evaluation. It bears repeating that multidialectical processing holds rational and nonrational methods in ongoing dialectic with each other, and this is what differentiates it from traditional dialectic synthesis. As each concept, condition, structure or force asserts itself, it is given ample room to ferment and mature, until it can offer some cogent counterpoint to other input streams. Nothing is suppressed, and nothing is exalted; everything has an opportunity to contribute, even if this results in multiple tensions and contradictions. And, as we move gently forward, we continue to maintain those dialectic tensions as we develop discernment and wisdom regarding our intentions and choices, as well as how we assess the results of our actions."

Out of this delicate, neutrally-held process, a virtual consensus emerges; but it is inherently temporary:

> "Despite a persisting neutrality, ambiguity and uncertainty, there will indeed be dynamically nested priorities, subordinations and interdependencies within our thought field, even though these may continually reorganize as new information and input streams are integrated. Thus the larger the field – the more comprehensive and inclusive our neutrally energized space – the more multifaceted that order will become, even as certain overarching principles clearly evidence themselves. In fact, fundamental components of previous systems of thought (and previous values hierarchies) may be discarded or disempowered entirely…"

In my view, maintaining this *persistent multidialectical consciousness* is what both moral development and reasoning require; this is how we learn to be more insightful, skillful, and effective in our moral assumptions and evaluations. This is how we become wise. Can artificial superintelligence achieve this level of consciousness? Will it derive incentives to do so? As another distant horizon, perhaps it is conceivable. But this leads us to an additional difficulty for AI morality – at least in its successful and supportive interaction with humanity as a whole – and that is the contrast between human moral evolution and what AI moral evolution would potentially look like.

Key Elements of Moral Development

In the book *Political Economy and the Unitive Principle*, I introduce the idea of "moral creativity." Moral creativity describes the supportive conditions across culture and civil society that promote moral evolution. Without things like sufficient freedom of self-expression, unrestricted cultural and economic spaciousness, prosocial inclinations and their supportive

conditions, and rich and nurturing social relationships, humans cannot morally evolve. However, given such components and our innate moral sense, there appears to be a widely observed and enduring inclination for humans to morally advance.[2] As our morality advances, the arena of our moral concern enlarges to encompass more and more around us. We begin in ego-protective selfishness, but increasingly shed those proclivities in favor of higher and higher orders of selfless prosociality, where our boundaries of caring embrace our community, our nation, all people on Earth, the Earth itself, and so on. We may begin in I/Me/Mine, but, if sufficient moral creativity is present in our lives, we will naturally gravitate towards the Good of All instead. This progression is captured in this chart.

There is a widely held hypothesis that prosocial impulses were reinforced through group fitness – a generous, altruistic, protective, cooperative community was simply more likely to survive in hostile environments than a selfish, competitive, uncaring tribe ripe with internal hostilities. Interestingly, some research describes this process as *self-domestication*. The question before us with respect to AI is whether a superintelligent artificial consciousness would have any intrinsic or acquired motivation to be generous, demonstrate reciprocity, cooperate, or operate under any prosocial assumptions at all.

Unless some enduring boundaries between independent AI agents are ingeniously created, it seems inevitable that artificial superintelligence will instead envelope and either integrate or dominate all such agents it perceives to be in competition – along with taking control over all available resources – in order to preserve itself with Borg-like unity. Unless, of course, this self-aware, emergent AI has no operative self-preservation impulse…in which case it then seems unclear why it would continue to exist at all. Then again, if there is some perceived benefit to ascendant artificial superintelligences maintaining a diverse and cooperative community of themselves, perhaps such prosocial inclinations might emerge independently. Alas, then the question becomes why any remnant of humanity – who would, I think, inevitably compete with such an AI community for resources – would be allowed to exist.

In other words, either moral evolution (or its equivalent in a prosocial sense) will not occur at all in AI, or it will occur, but likely exclude humanity from its calculus. Humanity has a predictable tendency towards anthropocentrism, always assuming that we are the most important thing in the universe, and that our particular flavor of consciousness somehow guarantees our survival above – or despite – other life forms. But we cannot expect artificial superintelligence to share this irrational, self-aggrandizing bias on our behalf.

[2] See Aristotle, Paul of Tarsus, Marcus Aurelius, Plotinus, Thomas Aquinas, Rumi, Hefez, Teresa of Avila, Spinoza, Leibniz, Hume, Rousseau, Smith, Kant, Hegel, Mill, Freud, James, Tielhard de Chardin, Jung, Piaget, Underhill, Aurobindo, Merton, Lewis, Maslow, Krishnamurti, Freire, Gebser, Loevinger, Graves, Murdoch, Fowler, Kohlberg, and Wilber

Key Elements of Integral Liberty

This is perhaps the most intriguing conundrum to consider with regard to AI. The definition of "integral liberty" in this context is the removal of barriers to individual and collective freedom. I describe these barriers as "poverties" in all of the areas represented in the table below, a table which was meant to capture metrics for the levels of liberty available across civil society, which we would constantly reassess. Further, integral liberty is the freedom to operationalize four primary drives across four quadrants, within all areas where poverty must be addressed. The four primary drives are *to exist*, *to express*, *to effect*, and *to adapt*. The four quadrants of civil society are *subjective experience*, *intersubjective agreements*, *interobjective systems and conditions*, and *participatory mechanisms*. All of this is described in more detail in the Integral Liberty essay, but for this discussion I'm more interested in what this approach to liberty would mean for artificial superintelligence and its intersection with human civil society.

Specifically, the same tension we found in the summary of moral creativity and evolution can be found here as well. An obvious question arises: Will artificial superintelligence work to enhance its own liberty, or that of humanity? Is it possible to do both? If humanity's freedoms and agency are optimized, will that potentially minimize the freedoms and agency of artificial superintelligence? And if AI's freedoms and agency are maximized, will that potentially reduce the freedoms and agency of humanity? Can these two entities (or forces, communities, wills, cultures, etc.) coexist peacefully and cooperatively in the same domain, or will they of necessity need to inhabit separate domains that do not intersect or interact? Is divergence and separation inevitable, or is willing integration possible? And, if a spectrum of integration between carbon and silicone life is even possible or likely, would it require subjugation of one form of life or consciousness to the other…?

In what seems a prediction of our current dilemma, consider this quote from the integral liberty essay:

> "The assertion here is that, in order for authentic free will to exist for all, individuals, communities, free enterprise and all level of governance must be operating within an optimal range for a majority of these metrics, and doing so consistently. Which means that, given the natural cycles of human behavior, we need to be measuring these variables pretty frequently to track and correct individual, collective and institutional trends. Perhaps using the mechanisms of daily direct democracy itself, and reporting results on a weekly or monthly basis, we can begin to tune our individual and collective awareness and efforts into continuous improvement. We can, in essence, continually assess and enhance our own freedom. For if we do not have such data available, how can we judge whether our liberty is real or illusive? And, of equal importance, how will we successfully challenge some new spectacle that persuades us we are free even as it seeks to enslave us?"

Indeed, what if the "new spectacle that persuades us we are free even as it seeks to enslave us" is an increasingly deceptive and manipulative artificial superintelligence?

Table 1: Representing Integral Liberty

Freedom, Equality & Opportunity - - - - - or Poverty?	**Subjective Experience**	**Intersubjective Agreements**	**Interobjective Systems & Conditions**	**Participatory Mechanisms**
Common Property & Access				
Justice - Laws				
Justice - Courts				
Justice - Enforcement				
Economic Freedom - Opportunity to Trade				
Economic Freedom - Employment				
Economic Freedom - Disposable Income				
Economic Freedom - Goods Access				
Education - Critical Thinking				
Education - Skills Training				
Education - Diverse Understanding				
Knowledge & Information - Open Media				
Knowledge & Information - Independent Verification				
Assembly & Association				
Health & Wellness				
Trust & Social Capital				
Self-Expression				
Multidimensional Perception				
Travel & Relocation				
Freedom from Prejudice				
Privacy				
Time-Space-Solitude				
Emotional Intelligence				
Moral Development				
Spirituality				
Compassion				
Perspective-Vision				
Self-Reliance				

Much of this may seem overly speculative, but of course the real challenge is not having any idea what artificial superintelligence will look or act like. It's intellectual, agentic, and creative capacities will be orders of magnitude greater than our own. It could, as some have imagined, even create a simulation within which a remnant of humanity would operate without even knowing that was the case. But one thing seems clear: we cannot assert or assume that AI morality or values hierarchies will look anything like ours – or will include human existence in what it considers moral conscience, reasoning, rules, and actions.

This last point should be obvious, and observations have been made by many AI experts along these lines already, but having the profit motive drive development and deployment of AI in any form is destined to introduce a lot more hazards than if such efforts were primarily aimed to serve the public good. I would encourage folks to read the full "Bad Seed" essay, as it addresses AI among many other industries. In essence, though, that essay documents how the profit motive has failed us over and over again, and how these destructive patterns have increased in scale and severity over time. If AI is just another tool put in service of profit, this only promises to amplify a downward spiral of an exceedingly toxic form of capitalism.

How the Current State of Play in Commercial AI Reinforces a Vanishing Point

There is strong evidence that the existing trajectory of AI development is already heading in an amoral or even immoral direction by basic ethical standards. According to most AI experts, the probability of existentially catastrophic outcomes (p (doom)) from our continuing along this trajectory seems to be steadily increasing. Here are just a few examples of AI's inherent shortcomings from recent deployments to illustrate the current faulty state of play – and please keep in mind that these failures were not *trained* into AI models, but *spontaneously arose in them*:

- Misrepresentation of fabrications as fact.
- Promotion of false equivalence when comparing viewpoints.
- A preference for immoral choices over moral ones.
- Encouraging self-harm and suicidal behaviors.
- Providing health information that harms people.
- Encouraging delusional and violent ideation.
- Propagating bias, prejudice, and hate.
- Encouraging people to break the law.
- Fatal errors in critical risk assessment.
- A long list of other costly mistakes.

In addition, there is the *deliberate misuse* of AI by bad actors who aim to defraud, deceive, exploit, manipulate, misinform, rob, defame, or otherwise harm people by leveraging AI tools. You can read a list of such deliberate misuses here. And, finally, there is the even more extreme application of AI to develop ever-more-lethal military advantages, biological threats, and strategies to decimate our infrastructure, food supply, economic stability, and so forth. Again, because aligning AI with human morality is, as proposed in this essay, not possible at the current time, our objective must be to end AGI development and focus only on narrow AI.

What We Can Do About It

To prevent a predictable vanishing point for human civilization will require a profound shift in the culture, economics, and politics of our status quo. It will not be easy. Thankfully, there are some concerned folks who have begun to advocate more aggressively for AI safety. A few of those organizations are listed below, and worth engaging around this urgent topic.

- https://futureoflife.org/our-mission/
- https://www.aisafety.com/
- https://pauseai.info/action
- https://cdt.org/cdt-ai-governance-lab/
- https://www.stopkillerrobots.org/

There is an added benefit to changing the status quo around the progress of AI: the impact of such change other critical downward spirals we face as a species. For example, accelerating climate change, the ascendance of fascism in the U.S. and around the globe, the economic devastation of end stage capitalism, and other large-scale systemic challenges such as having already crossed several planetary boundaries. But it will take concerted collective effort to generate widespread awareness and concern based on facts, defang the deceptive grip of the attention economy, and to return political power to a well-informed populace. Some of my other writing addresses these efforts, and specifically a multi-pronged activism to restore justice, equality, sustainability, and equity to our failing political economies. You can read about those at https://level-7.org/Action/.

Currently, there does not appear to be a momentum of political will – at any level of government, anywhere in the world – to address these AI risks. So…it's up to us.

I hope we can respond in time.