The Maximally Distributed Intelligence Explosion

Francesco Albert Bosco Cortese

Affiliate Scholar of the Institute for Ethics & Emerging Technologies; Research Scientist at ELPIs Foundation for Indefinite Lifespans;
Assistant Editor of Ria University Press; Chief Operating Officer of the Center for Interdisciplinary Philosophic Studies
francocortese1@gmail.com

Abstract

We argue that the most effective solution paradigm in machine ethics aiming to maximize safe relations between humans and recursively self-improving AI is to maximize the approximate equality between humans and AGI. We subsequently argue that embedding the concomitant intelligence-amplification of biological humans as a necessary intermediary goal between each successive iteration of recursive self-improvement – such that the AGI conceives of such an intermediary step as a necessary sub-goal of its own self-modification – constitutes the best logistical method of maintaining approximate intelligence equality amongst biological humans and recursively self-improving AGI. We ultimately argue that this approach bypasses the seeming impasse of needing to design, develop and articulate a motivational system possessing a top-level utility function that doesn't decay over repeated iterations of recursive self-improvement in order to have a safe recursively self-modifying AGI.

Superintelligence: A Double-Edged Cliff

Superintelligence is a sharper double-edged sword than any other. It constitutes at once the greatest conceivable source of existential risk and global catastrophic risk, and our most promising means of mitigating such risk. Superintelligence possesses at once greater destructive potential than any other emerging technology and greater potential to formulate new solutions to emerging existential risks and global catastrophic risk because it is the very embodiment of the ability to conceive of new weapons and new solutions to existing threats. A superintelligence could not only thwart any prior-existing security measures against emerging-technology-mediated existential risk, but it could also formulate new generations of weapons so superior as to be inconceivable to those of lesser intelligence.

This poses a grave problem, as developments in artificial intelligence, combined with continuing increases in computational price performance, are making the creation of a superintelligent AI (or more accurately a recursively self-modifying Seed AI able to bootstrap itself into superintelligence) easier and cheaper. Malicious or merely indifferent superintelligent AI constitute a pressing existential risk for humanity.

This has motivated attempts to formulate a means of preventing the creation of a malicious or indifferent superintelligent AI. The solution paradigm that has thus far received the most attention is 'Friendly AI', or more accurately the notion of a 'Coherent Extrapolated Volition Engine' (a.k.a. CEV) as formulated by the Machine Intelligence Research institute (a.k.a. MIRI, formerly the Singularity Institute). A CEV would be a recursively self-modifying optimization algorithm whose topmost utility function does not decay over recursive iterations of self-modification, and whose topmost utility function is designed in such a way that the CEV formulates what humanity would desire if we were smarter and more amenable to consensus.

The construction of CEV is motivated by the concern that the first superintelligent AI will be built in a way that makes no attempt to ensure its safety of 'friendliness' relative to humanity. This is indeed an important and pressing concern. But I and others argue that MIRI is going about it in a fundamentally misguided way. They seek to prevent the creation of a rogue superintelligence by creating the first one, and making sure it's built as safely as the technology and methodology of the times can manage. This is somewhat akin to trying to prevent the creation of nuclear arms by being the first to create it and then connecting it to a global surveillance system, and using that nuclear weapon to threaten anyone else who might be found to be building one.

Admittedly, superintelligence has some properties which make this line of attack – i.e., being the first to create superintelligent AI and building it to be as safe as one can – seem appealing and intuitive. For instance, upon the creation of a superintelligence, the rate at which that superintelligence could gain control (defined here as the capacity to effect change in the world, and approximately analogized with intelligence, such that a given degree of intelligence would correlate with an agent's degree of control, i.e., the agent's capacity to affect changes in the world) is unprecedented. This means that upon the creation of an effective Seed AI, the battle is largely already lost. So being the first to make it would in this case constitute a palpable and definitive advantage.

However, we argue that a number of alternative solution paradigms to the existential risk and global catastrophic risk posed by the creation of a malicious or indifferent Seed AI exist and warrant being explored as alternatives to CEV. Whereas many prior solution paradigms sought to minimize the unpredictability of a superintelligence, we argue that unpredictability is inextricably wed to the property of superintelligence, that it is one of its most definitive and essential characteristics, and that to remove unpredictability from superintelligence is to remove superintelligence itself. The whole point of creating such Seed AI I to think and do that which we as humans cannot, to think thoughts that are categorically unavailable to us as intelligent agents. To seek this while simultaneously seeking a comprehensive and predictively-accurate understanding of those as-yet-unconceived-of products that the AI is meant to bring into being is tautological.

Most critics of CEV have challenged it on the grounds of feasibility, arguing that a recursively self-improving optimization algorithm possessing utility functions that remain stable over recursive iterations of self-improvement We agree, as expressed above, but also advance two ethical concerns over its development that further bring into question its merit as an effective solution paradigm to the existential and global catastrophic risks posed by Seed AI. We argue firstly that it even if its feasibility weren't in question – i.e. if a Seed AI that is predictable, which 'friendly' utility functions that do not decay over iterations of recursive self-modification, were definitively feasible) – then it would still be unethical to create any intelligent agent that decides the fate of humanity for them, at whatever scale and in whatever context. The heart of the human is our will toward self-determination – our unerring attempt to determine the conditions and circumstances of our own selves and lives. It is exemplified by the prominence of autonomy and liberty, or inviolable human rights and democracy, as human values throughout history. It informs and is infused throughout almost all that we do as intelligent agents. Articulating one's complex goals in complex environments is an instance of self-determination and autonomy. And to create any entity that decides the determining circumstances of humanity on the level of the collective or the individual is unethical because it is dehumanizing and an affront to our very essence as fledgingly self-determining creatures. To argue that it is in humanity's best interest to vest all control over their lives and circumstances into any single intelligent agent is undemocratic and contrary to both universal human values (autonomy and liberty) and our most definitive essence, our longing to have more control over the determining conditions of our own selves and lives. Secondly, we argue that CEV and the solution paradigms it exemplifies would also be unethical for a different reason – namely that restricting any intelligent, self-modifying agent to a specific set of values, beliefs or goals – i.e. a preprogrammed and non-decaying utility function – is unethical on the same grounds – namely that externally determining the determining conditions of any entity possessing some degree of self-modification and self-determination (i.e. any self-modifying intelligent agent) is unethical because it is directly contrary to their foremost values (autonomy and liberty) and to their coremost essence, i.e. their longing to determine for themselves the conditions and circumstances of their own lives and selves. For these reasons we argue that it would be unethical to create a Seed AI without also allowing it to formulate its own ethical system, in accordance with its own self-formed and ever-reformulating beliefs and desires. Moreover, to restrict such a Seed AI to the moral codes and beliefs of a kind of entity (human) that it was built for the express purpose of surpassing in thought and intelligence is even more unethical and for the same reasons.

Contrary to such past solution paradigms, we argue that the existential risks posed by any single entity with a level of intelligence (and thus control, as defined above) significantly surpassing other entities it has the capacity to interact with far exceeds the potential advantages offered by its creation – such as new solutions to humanity's gravest crises and concerns, like disease, poverty and pollution. Furthermore, we articulate a new solution paradigm for mitigating the existential and global catastrophic risks incurred by the creation of a recursively self-modifying seed AI, the end goal of which is not a safe superintelligence, but rather the amplification of intelligence without ever incurring the relative superintelligence of any agent over another. In other words, it seeks to facilitate a maximally-distributed intelligence explosion, aiming to maintain rough equality of intelligence (and thus control) amongst all intelligent interacting agents. The only way to have safe superintelligence is to do it globally – in which case no one is superintelligent relative to anyone else, but rather only superintelligent in relation to those who came before – or ideally, only superintelligent relative to past instances of one's own self.

Most critics of CEV have challenged it on the grounds of feasibility, arguing that a recursively self-improving optimization algorithm possessing utility functions that remain stable over recursive iterations of self-improvement

International Legislation on Seed AI

Another possible solution paradigm for mitigating the risks of Seed AI is to characterize it as a potential Weapon of Mass Destruction (which is entirely conceivable considering its potential military applications) and to develop international legislature regulation its development and use similar to the international regulation of nuclear arms and other weapons of mass destruction.

While we do not have a historical precedent for international legislation on AI, we do have a number of solid historical precedents for developing international regulations around emerging technologies with vast destructive capabilities, like nuclear arms and biological weapons. Furthermore, the motivation for developing such regulations is the same as the motivation for regulations on WMDs – namely their vast destructive capabilities.

This solution paradigm is less effective for Seed AI than it would be for WDMs, largely because the development of a Seed AI could put its developers in a vastly superior defensive and offensive position, and as such a much better position to ward off any repercussions resulting from the transgression of such international law. It nonetheless serves as a potentially fruitful alternative to other solution paradigms for mitigating the risks of Seed AI, like CEV. This solution paradigm bears similarities to another solution paradigm called "Global AI Nanny", which would seek to implement a non-superintelligent AI connected to a global surveillance system aiming to stop anyone it catches working on Seed AI. Furthermore, the use of such a solution paradigm in conjunction with international regulations on Seed AI could help deter the disadvantage described above – namely the fact that once a Seed AI is created, the capacity of their developers to escape penal repercussions could be vastly increased, depending on the Seed AI's rate of self-modification. This would allow such developers to be caught before they succeeded in fully developing their Seed AI, thereby obviating the seeming dilemma.

The Maximally Distributed Intelligence Explosion

The present solution paradigm seeks to infuse approximate equality of intelligence into a given Seed AI's self-modificational infrastructure itself.

A Seed AI will have to view the problem of increasing the intelligence of humans as just one of the series of problems it needs to solve in order to implement its next iteration of self-modification. After all, the universe or the design constraints and parameters of its first implementation (which it didn't have any control over ipso facto) could have been different in such a way as to make the number of activities it needs to articulate to perform the next iteration of self-improvement many more than they are now; why should the need to modify humans so as to reach approximate equality with their next target-intelligence-level not also constitute a sub-problem along the way to the AI's next increment of intelligence-amplification.

A Seed AI breaks its intelligence amplification into increments. At each increment it must develop an intelligence-equivalency test (much like the Turing Test) to determine equality of intelligence between itself and humans, and then conceive of a means of implementing intelligence amplification in biological human nervous systems such that the intelligence of the agents possessing those nervous systems is approximately equivalent to the Seed AI at that increment of intelligence amplification, as judged by the intelligence-equivalency test. This could involve the use of neurotechnology, biotechnology and/or nanotechnology so as to implement physical changes in biological nervous systems that lead to an increase in intelligence. By making such parallel improvements a fundamental step within the larger problem to advancing to the next increment of intelligence amplification, by making it a step that the AI seeds as a step toward its own self-modification, is a possible means of implementing a maximally-distributed intelligence amplification within the context of a single (or group of) Seed AI.

Factors Making the Ease and Rate of Self-Modification Greater in Seed AI than Humans

There are several factors biasing Seed AI (or AI in general) toward a greater ease of self-modification and a greater potential rate of self-modification than biological humans. This is disadvantageous, as the aim of the solution paradigm articulated in the present paper is to maintain rough equality amongst the rate of self-modification among intelligent interacting agents.

For instance, in biological humans a rate of self-modification that is too high could result in phenomenal discontuity – that is, discontinuity between past and future instances of oneself. The physical brain is changing its physical organization all the time, but it is doing so gradually. If we were to reorganize the synaptic connection in our brain too fast, or if we were to reorganize too many synaptic connections in our brains at once, could result in phenomenal discontinuity just as taking out a large portion of my cortex and replacing it with another person's might result in phenomenal discontinuity. Thus the rate of self-modification in biological humans is limited by the rate at which they can self-modify without incurring phenomenal discontinuity.

AI, by contrast, would not necessarily have any motivation to avoid a rate of self-modification so high as to incur phenomenal discontinuity. Evolution has not engrained it with a fear of death and an instinctual desire to possess phenomenal continuity.

Another example of the biases AI has toward greater ease of self-modification and a greater rate of self-modification lie in the fact that what it seeks to modify is software rather than hardware. In a biological nervous system, in order to self-modify one needs to determine not only what modifications will result in intelligence amplification (e.g. increased synaptic density), but what methods and technologies are needed to physically implement those changes (e.g. nanomedical systems to implement changes in the physical organization or structure of neurons). With a Seed AI, because it is software, the implementation side of things is already solved. Modifying itself is as easy as rewriting words on a file. This serves to increase a given AI's potential ease of self-modification, and thereby it's potential rate of self-modification.

Thus formulating ways to balance the ease of self-modification and potential rate of safe self-modification in humans and in AGI would constitute progress towards the goal of a maximally distributed intelligence explosion.

Does this Negate the Utility of Seed AI?

On the level of aim and motivation rather than implementation, making everyone superintelligent might seem to negate the utility of having a single superintelligent agent in the first place: namely the existence of a single agent more intelligent than humanity, so that we don't have to do the hard work ourselves. But this notion mistakes the end for the means and the motivation for the mediation. The entire point is to have the means of remediating some of humanity's foremost concerns and crises, i.e. the intelligence to improve the state of the world and its inhabitants. The historical context in which this notion came into being was one that seemed to preclude self-modification as a means toward that end. It didn't occur to us then that the creation of a more intelligent, more self-determining entity might take the form of a reformatted and/or reformulated self, rather than a newly-formed other. Humanity has been trying to recreate mind in the media of mindless matter long before the inception of Artificial Intelligence proper. Artificial intelligence has its historical antecedent in the computer, which sought to implement and automate systems of logic first formulated by humans – and long thought of as the basis of mind and the heart of our human essence – in non-biological systems. And even before Babbage, the computer too has its historical antecedent in the creation of mechanical calculators that sought to implement that highest of human faculties, mathematical reasoning, abstraction and symbolic manipulation, in the medium of mechanical gears and cogs. It is important to note that the first mechanical calculators were created within the context of the growing Enlightenment tradition, which heralded human reason as the noblest human faculty and located it as the very basis for and fundament of our autonomy and liberty. Human reason and autonomy are hallmarks of the Enlightenment tradition, and the fact that the first mechanical calculators arose during such a time says much about the grandiose ontological stakes at play in the seemingly-mundane – or at most practical but decidedly non-metaphysical – creation of the mechanical calculator. But seen in the historical context in which it arose, the development of the mechanical calculator is nothing less than an attempt to imbue soulless, mindless matter with what was then considered the highest, noblest, rarest and most distinguishing human faculty: abstract and mathematical thinking, which was at the time considered the very core of our reason, autonomy and liberty.

The impetus to create artificial intelligence is intimately bound to our deepest human values and concerns – our own mortality. We work to imbue matter with mind as a minute tribute to death's funeral pyre. We instill nothing less than life itself – and its higher-order counterpart mind – into collections of disparate components as a chaste transgression of our own deathly dissolution and a cherished blasphemy against our own oncoming demise. If life and mind can be created by the hand of man – if life can be assembled and created as naturally as it can be disassembled and destroyed – then the vast metaphysical mystery behind death is made a bit smaller and a little less real.

The impetus to create artificial intelligence is also intimately intertwined with another core human longing and value – growth and transcension of the self. Most of us in one way or another seeks to work on or contribute to causes and goals that transcend our own lives and circumstance. This too is a revolt against our mortality. By contributing to large-scale endeavors that will transcend our own lifetimes, we are imparting indirect causal influence upon the future state of the world. We are in an indirect but very real sense affecting the future and imparting actions that will leave a mark on the world after our own marks have long since washed away. Thus the impetus to create artificial intelligence is also intimately informed by our very human longing to create the godly and to contribute to something extending beyond the bounds of our own lives and circumstances. This is what makes Hugo de Garis characterize the creation of superintelligent AI as a deeply profound, spiritual and even religious practice for many.

Thus man has been working on the recreation of the newest and thus truest form of man, in the frame of another, for a very long time. Indeed, stories like the Golem in the Jewish religious tradition, which depicts a previously-lifeless clay entity being animated by nothing less than an abstract coded language – serve to exemplify how close to humanity's heart the recreation of ourself really is.

But the notion that we might recreate ourselves rather than recreating our self – and i.e. the recreation of our individual minds rather than the recreation of mind in general within the frame of another, i.e. an entity external to the self – seems to be a relatively new notion in the context of human history. For much of human history we thought of the self as a metaphysically-insulated, singular and static entity. It hasn't until the rise of scientific materialism in general and modern medicine in particular that we began to conceive of the self as emerging from the physical brain, a system of connected components that are dynamic rather than static and decidedly non-singular. Instances and experiments wherein physical modifications made to the brain resulted in changes to our phenomenal consciousness gradually taught us that changing one's mind was more than metaphor, and that the manipulation of the physical parameters of our brains could allow us to effect targeted changes to our own phenomenal consciousness. Such realizations laid the groundwork for the modern notion of self-determination as self-modification.

0 The notion that a maximally distributed intelligence explosion would negate one of the main utilities of a superintelligent agent – namely of there being a single agent more intelligent than us to solve humanity's problems, so that we don't have to do the hard work ourselves – again, misses the means for the motivation.

The notion of mind uploading (i.e. constructing a predictively-accurate computational simulation or emulation of a mind and then 'transferring' such a software mind to a computational substrate) and later of gradual mind uploading (i.e. replacing the constitutive components of the brain with computational simulations one component at a time, so as to preserve phenomenal continuity or the subjective perception of being phenomenally-continuous with past and future instances of oneself) brought this notion further. If it were conceivable to gradually replace our physical brains with a simulated counterpart, and to maintain our own phenomenal continuity throughout the procedure, then our capacity to self-modify – our degree of self-modifiability – could be vastly increased. The notion of mind uploading later became the historical antecedent of the modern discipline of Whole Brain Emulation.

One of the main utilities of recursively self-modifying AI lies in the fact that its mind is software rather than hardware. This presumably would allow such an AI to directly rewrite its own source code. This makes the implementation of self-modification categorically easier. In order to implement intelligence-amplifying modifications to a physical nervous system one needs to know not only what changes to make, but how to actually physically implement them. We need to formulate (a) what changes would effect an increase in intelligence, and then (b) construct methods and technologies for physically articulating the changes that would result in (a). But when the system one is looking to modify is software rather than hardware, the implementation side of self-modification is made as easy as rewriting data on a file.

Thus the solution paradigm suggested in the concluding section of this paper might seem to negate the very utility of a recursively self-modifying superintelligence in the first place. But note that through the notion of 'gradual mind uploading', the implementation of human self-modification can likewise be made categorically easier. As software (i.e. the simulated analog of the brain's physical components and their integral operation), effecting changes that would have previously (i.e., in a physical nervous system) required a whole host of new methods and technologies to implement becomes as easy as rewriting data on a file. Thus if this route to human self-modification is taken, the utility of a vastly increased degree of self-modifiability does not remain an exclusive feature of AI, and can likewise be a property possessed by self-modifying humans as well.

Today, the growth and progression of man has two potential media – recreation of ourselves in the formation of another, and recreation of ourselves through the reformulation of ourselves, i.e. through self-modification and self-determination. Whereas historically we sought to articulate our mutiny against mortality and our longing for growth the only way we knew how, i.e. the recreation of another, today we have two choices. And it is the latter of these two that is the safer, the more ethical and the more effective option.

The present solution paradigm is safer because it allows approximate equality of intelligence, and thus of control (again, defined here as capacity to affect changes in the world), to be maintained amongst intelligent interacting agents. Furthermore, the present solution paradigm allows for the gradual amplification of intelligence (and all its benefits and utilities that today partly motivate the development of Seed AI) without actually incurring the relative superintelligence of any one agent over any other. This allows us to reap the benefits of increased intelligence without actually incurring the aspect that poses the greatest existential and global catastrophic risk, namely relative superintelligence of any one agent over another.

Secondly, it is more ethical because it does not necessitate that we impose our beliefs or values upon any other intelligent agent or self-modifying agent, or that we explicitly restrict the development of any agent’s ethical system beyond the level of an organism that it is built so as to surpass in capability and intelligence. It is also more ethical because it is safer, as implementing or creating anything that poses grave threats greatly outweighing their potential benefits can be seen as unethical.

Thirdly, it constitutes a more effective method of producing results (e.g. thoughts, solutions to existing problems facing humanity, etc.), in the sense of our confidence in the accuracy or competence of the answers a superintelligent agent produces. This is simply a result of the fact that any solutions will be arrived at by a number of separate agents independently, rather than by a single entity with its own biases. The fact that any answer agreed upon would be arrived at independently, and furthermore the fact that such answers would then be debated amongst multiple intelligent agents, serves to reinforce the solution paradigm’s effectiveness at producing accurate and competent answers. An analog of this situation can be seen in biology, wherein genetic diversity provides benefits. If we take all our resources and apply it to one possible solution to a problem, if we’re wrong then those resources were wasted for naught. But if we distribute such resources amongst several competing possible solutions, the probability of one of the yielding an accurate or competent answer is increased.

Thus the notion that one of the main utilities of a superintelligence is the fact of it being a singular entity with relative superintelligence in comparison to agents (like humans) is, we contend, an outdated memetic child of a time when the continuation of the self and contribution to something more grand than our own lives and circumstances seemed more likely to come from the recreation of mind anew in a new form and frame, as a separate entity, rather than from the recreation of our own selves, through a gradual process of recursive self-modification. It is the product of a time when the continuation of the self was best facilitated by the rearing of children, on whom parents often impress their own ideals and worldviews, making their children not only a continuation of themselves in terms of body and biology, but in terms of mind as well. It made complete sense to turn to the creation of another in our dire efforts towards the recreation of our own selves. But it makes sense no longer, because a safer, more ethical and more effective alternative has since come into suggestion, i.e. recursive self-modification rather than the creation of a de-novo recursively self-modifying entity.