Conceptual-Linguistic Superintelligence

David J. Jilk
eCortex, Inc., 9035 Wadsworth Pkwy, Suite 2275, Westminster, CO 80021-8675, UK
E-mail: dave@jilk.com

Keywords: artificial intelligence, superintelligence, intelligence explosion, existential risk

Received: August 31, 2017

We argue that artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have a conceptual-linguistic faculty with substantial functional similarity to the human faculty. We then argue for three subsidiary claims: first, that detecting the presence of such a faculty will be an important indicator of imminent superintelligence; second, that such a superintelligence will, in creating further increases in intelligence, both face and consider the same sorts of existential risks that humans face today; third, that such a superintelligence is likely to assess and question its own values, purposes, and drives.

Povzetek: V prispevku je predstavljena teza, da je za superinteligenco potrebna tudi konceptualno-lingvistična inteligenca, ki mora biti vsaj delno podobna človeški.

1 Introduction

Recently much analysis and speculation has been offered to describe scenarios related to a possible intelligence explosion, a notion first suggested by I.J. Good [1]. In an intelligence explosion, initial creation of artificial intelligence with a critical mass of capabilities and drives is followed by an inexorable process of increases in that intelligence. Eventually the resultant artificial intelligence exceeds human intelligence and is referred to as superintelligence. This process is usually viewed as uncontrolled, unstoppable, and accelerating; the scenarios have generated considerable consternation and are driving a conversation about a number of ethical and technological issues [2] [3] [4].

In this paper, we argue that artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have a conceptual-linguistic faculty with substantial functional similarity to the human faculty. We follow this with arguments for three subsidiary claims: first, that detecting the presence of such a faculty will be an important indicator of imminent superintelligence; second, that such a superintelligence will, in creating further increases in intelligence, both face and consider the same sorts of existential risks that humans face today; third, that such a superintelligence is likely to assess and question its own values, purposes, and drives.

These conclusions do not guarantee a satisfactory outcome for humans, but do suggest that the process will be subject to ongoing scrutiny by its own participants. We note that it is possible that superintelligence may be created outside the context of an intelligence explosion; for example, humans might create it directly. Our arguments are not intended to apply in that case: we do not argue that a conceptual-linguistic faculty is required to constitute superintelligence (though this may be true), only that it is required in an intelligence explosion. Also, there are many risks of artificial intelligence aside from superintelligence and intelligence explosions, such as those arising from autonomous weapons and unexplainable decision processes. We do not address those issues at all. Nevertheless, the existential risks associated with an intelligence explosion are an important topic in artificial intelligence safety [5] and we will hopefully deepen our understanding of those scenarios through this analysis.

2 The need for a conceptual-linguistic faculty

In this major section we begin by outlining implied necessary conditions for an intelligence explosion, and characterize in some detail what we mean by a conceptual-linguistic faculty. With that in place, we argue the foundational claim that a conceptual-linguistic faculty is necessary for an intelligence explosion to be sustained. We close the section by showing how the presence of a conceptual-linguistic faculty is a harbinger of superintelligence, and discuss how it might be detected.

2.1 Requirements for an uncontrolled intelligence explosion

As it is typically envisioned, an intelligence explosion comprises a sequence or continuum of artificial intelligence systems with progressively increasing intelligence. We will refer to each of these systems as a “participant.” Since a participant is part of a sequence, it has predecessors and successors in the process, with humans as the initial predecessor. In progressing the sequence, a participant may elect to self-improve or to create a new system; in either case we refer to the resulting system as a successor.

There are many factors to consider in assessing whether an intelligence explosion is likely to occur and how rapidly it might proceed [3] [6]. In this paper, we intend to focus on the role of the participants, and generally assume that extrinsic factors (for example, technical or resource recalcitrance) are favorable for supporting an intelligence explosion. Our emphasis will be qualitative and directional rather than quantitative.

For an uncontrolled intelligence explosion to occur, the progress of intelligence increases must be *self-sustaining* and *resistant to premature termination.* Though it is an analogy only, these requirements are similar to those of a nuclear fission weapon [4]. Two of the most difficult challenges faced by the initial designers of such weapons were to have a sufficient fraction of emitted neutrons be absorbed by fissile nuclei (self-sustaining), and for the chain reaction to proceed sufficiently before its own energy caused dispersion of the fissile material (premature termination) [7].

Corresponding requirements in an intelligence explosion are that each participant artificial intelligence has, as necessary but not sufficient conditions, these properties:

1. *Self-sustaining:* the participant must aim to and be capable of designing and building either self-modifications or new systems that have greater intelligence, without assistance from humans;

2. *Resistant to premature termination:* the participant must be capable of preventing other agencies, such as humans or later predecessors, from interrupting the development of self-modifications or new systems with greater intelligence.

An uncontrolled chain reaction is only worrisome if it produces *side-effects*. In the case of nuclear fission, each fission releases energy that contributes to the explosion. In the case of an intelligence explosion, the side-effects arise from the *goals* or *purposes* of the artificial intelligence. These purposes are potentially problematic for humans whether or not humans attempt to stand in the way.

2.2 What is a conceptual-linguistic faculty?

Humans evidently have the ability to organize their experiences into concepts, and to use language to access those concepts and thereby refer to aspects of those experiences. We will use the term "conceptual-linguistic faculty" to refer to this capability, whether possessed by a human or an artificial intelligence. There are numerous theories and considerable empirical insight about the mechanisms involved in concept formation and use, though we still do not fully understand them. However, for our purposes it is only necessary to gain some purchase on the kinds of functions it performs, particularly since the implementation in an artificial intelligence may be very different.

The centerpiece of the conceptual-linguistic faculty is the way it combines information representations that are treated discretely or symbolically with information representations that are graded, statistical, and overlapping [8] [9]. The former we will call "words" and the latter we will call "semantic contents." Semantic contents develop through perceptual-motor experience, and are activated by perceptual stimuli that are sufficiently "similar" [10]. "Activated" means that the representations temporarily obtain some sort of facilitated access and priority of influence in current cognitive processing; activation is graded rather than binary [11]. What constitutes sufficient similarity is embedded in the semantic contents themselves and can be extremely complex and multi-dimensional.

Some semantic contents are bi-directionally attached to a word, and we will call such a pair a concept. The nature of this attachment is that when the semantic contents are activated by a stimulus, the word is also activated; and when a word is activated through memory or communication, even in the absence of applicable stimuli, the semantic contents are also activated [12]. However, this description suggests a crisp boundary, and it is nothing of the sort. The semantic contents activated by a stimulus depends on detailed features of the stimulus and the context; the word activated by the semantic contents depends on the context [8] [13], which may include attention. Further, the activation of semantic contents often causes the activation of multiple words at varying strengths [14].

Words also activate each other, and semantic contents can activate other semantic contents [14]. Semantic contents, which we have already mentioned do not have sharp boundaries, can be activated simultaneously, partially, and in many combinations, and simultaneous activations result in cross influences, sometimes called "dynamic realization." Crucially, we can use words and semantic contents to cognitively *simulate* the world and explore what the consequences of various actions or circumstances might be [13]. But we can also manipulate words as symbolic entities and process logical thoughts with minimal activation of the semantic contents, essentially treating the words themselves as objects [8] [15].

Despite this highly complex, graded, and overlapping network of relationships, the informational representations offered by concepts have sufficient structure and distinctness to enable humans to create models of the world, make successful predictions, design and build sophisticated tools, share experiences, and the like. Our description here relies on results in cognitive psychology and neuroscience; readers may note that our citations include some opposing theorists who nevertheless mostly agree that human cognition exhibits these basic features. Again, we do not know all the details of implementation of this capability in humans. What we do know is that other animals do not have it in sufficient quantity to build technological civilizations, and to date, no artificial intelligence has it.

An illustrative example may be helpful toward understanding what is meant by a conceptual-linguistic faculty; however, the account above and its associated references, and not this example, are the foundation of the arguments to follow. Suppose one is walking alongside a downtown street and perceives a building. The effects of this perception rely on having been previously exposed to many buildings that vary along many nonspecific dimensions, as well as many other objects that are not buildings. It triggers a passive, partial activation of the word “building” but more strongly the word “bank,” of which this building happens to be an instance. The word “bank” has a statistical connection to the word “mortgage” and the word “money,” which might now activate a visual representation of a mortgage statement or a stack of dollar bills; or it may activate one’s representations of money in general and its social and legal role, which may further activate the word “bitcoin.” Or the perception of the building may activate an olfactory representation of the inside of an old, marble bank lobby which further triggers representations of buildings in which sound echoes, which then activates the word “echo.” One might then entertain a relatively abstract linguistic thought such as “I will not be able to pay my mortgage without putting more money into my account,” or visually simulate logging in to the bank’s web site to effect the transfer. The likelihood of each of these derivative activations depends on current context and goals, among other things. Note the bidirectional interplay of the symbolic-linguistic representations with the perceptual-semantic representations, as well as interactions directly among percepts and directly between words; also note the graded, statistical character of all these interactions.

There may be many ways to implement a conceptual-linguistic faculty in artificial intelligence. Though a “neuromorphic” approach is an appealing candidate, since it has a reference implementation, it is not known to be a requirement. The key is that semantic contents of the sort we have described *refer* to the real world richly and bi-directionally, and *ground* the conceptual structure to reality.

Deep learning methods [16] illustrate the power and importance of rich grounding. In the past decade, these methods have demonstrated impressive success in classification, including both auditory and visual perception as well as more abstract patterns. The methods are mechanistically homologous to human semantic processing in several ways, ranging from learning rules to the hierarchical network structure and receptive field overlap at each layer. The statistical, graded, and overlapping representations afforded by such methods seem to be essential to their success. Still, though they may (or may not) represent a step toward successful general artificial intelligence, to date they lack important capabilities of a conceptual-linguistic faculty. Their “symbolic” representations (which might be likened to words) are impoverished and do not mutually interact, and they are not bidirectional while in operation.

Most past attempts to implement language and concepts in artificial intelligence are manifestly insufficient to produce a conceptual-linguistic faculty with the features we have described. In particular, historical efforts have often been purely symbolic in their representation of semantic contents. Systems like Cyc [17] or Prolog struggle to emulate human-like reasoning because they are entirely ungrounded – words or symbols interconnect with each other but have no means of referring to the world [18] [19]. Attempts to ground such systems via hardcoded algorithms to identify standard human conceptual classes (e.g., face and object detectors) provide limited grounding but entirely miss the complex overlap and subtlety of the real world [20], and are in any case feedforward and incapable of dynamic realization. Consequently, while some limited linguistic stimulus/response capabilities can be derived in these systems, they cannot really *use* the human concepts to do anything in the world, except in tightly constrained environments or simulations where the abstractions on which they rely can be simulated perfectly. Whether or not such approaches can be made to perform useful functions, they do not understand the meaning of human words and concepts in a way that enables them to perform flexible cognition with them. The deep reasons for these difficulties have been explained philosophically [21] [22] [23], and empirically [24].

Committed scientific realists are unlikely to be convinced by such arguments, for the same reason that they were surprised by Moravec’s Paradox, and continue to struggle with the “frame problem” [25] [26]. They hold that the world consists of objects that intrinsically belong to metaphysically distinct categories, thus all that is necessary for grounding is to find ways to identify those categories reliably. In contrast, our characterization of a conceptual-linguistic faculty relies on a view that it is cognition that constructs and ascribes categories. While the applicable terms refer to genuine clusters of features in the world, their categories are not the only way to partition experience, and they overlap in complex ways. The impressive success of deep learning has convinced many researchers of a need for sophisticated grounding, but there are holdouts. A thorough argument against scientific realism is beyond the scope of this paper.

2.3 A conceptual-linguistic faculty is necessary for an intelligence explosion

We now proceed to the foundational claim: artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have, at a minimum, a conceptual-linguistic faculty with substantial functional similarity to that of humans. We will argue this in two parts, corresponding to the two general requirements for an intelligence explosion. Importantly, this claim only asserts a minimum. We are not claiming that this faculty must be the only, or even the most effective or important, cognitive mechanism of a participant artificial intelligence. A participant may have many other capabilities which may be better at other things.

For the intelligence explosion to be self-sustaining, the participant artificial intelligence must be able to produce self-improvements or improved successors without human assistance. If it requires human assistance, then humans could withhold that assistance to terminate the process prematurely. If it is not capable of grasping and using human language and concepts, *including their underlying rich semantic contents*, then the vast body of human knowledge is not at its disposal.

Outside of human minds, the human body of knowledge is encapsulated primarily in words, diagrams, and functional artifacts (e.g., tools and machines). Words and diagrams rely on concepts; earlier, in discussing ungrounded and feedforward-only semantic representations, we elaborated why these cannot be used effectively without a conceptual-linguistic faculty. The purpose, means of use, and implementation of functional artifacts is severely underdetermined. Learning how and when to use such artifacts would require demonstration by humans or comprehension of instruction manuals. Understanding how they are built would require reverse engineering, which would be extremely difficult without a conceptual apparatus that could reproduce the conceptual model used for the original design [27].

We might wonder whether some sort of cognitive shortcut could be devised so that the conceptual knowledge embodied in language is translated into another form. A simple thought experiment illustrates why it cannot. Suppose that an artificial intelligence accesses all the equations of known physics and a thorough explanation of them. The speed of light figures prominently in these equations. What is light? We can provide experiential examples: the sun, lightning, fire, reflections. But these experiences need to be abstracted so that more than just the immediate examples provided can be used. The tools humans use to produce and detect light and measure its speed must be represented. The parts and components of those tools and their interrelationships must be identifiable in the face of noise and complex variation. To use such a tool requires visualization or simulation of its function. The explanation, written in human language, will use words intended to activate semantic contents and thereby (in a contextually appropriate way) activate other words and contents that enable comprehension. All of its words must refer in some way to entities in the real world, and identifying these entities requires a means of perception that is variation tolerant in a high number of dimensions. The existing literature varies tremendously in its precision and consistency; there is even a considerable body of poetry that invokes light, and these sources must be somehow interpreted and distinguished. Ignoring all these details and "hardcoding" the speed of light requires a brittle and limited pre-programming, and merely defers the difficulty to other, equally complex notions that regress indefinitely.

We can see that all of the attributes of the conceptual-linguistic faculty, as described earlier, are required to unpack the reference to the speed of light in our human equations and explanations and make use of it in novel ways in the real world. This is not a failure of imagination – it speaks to the essence of how human knowledge is constituted, and therefore how it must be understood.

In the development of technology, it is a distinct advantage merely to know that something is possible and can be made to work. Artificial intelligence that lacks a conceptual-linguistic faculty will not even be able to observe humans and their tools to gain that knowledge. Without grounded concepts that enable comprehension of either human words or even conceptually-structured observation of human activity, such a system will have no shortcuts to grasping what is technologically possible.

Importantly, we are not making the much stronger claim that an artificial intelligence must have a conceptual-linguistic faculty to understand and operate in the world. We are only claiming that such a faculty is necessary to understand and make use of the human body of knowledge. Nevertheless, this leads to a possibly startling conclusion: an artificial intelligence lacking a conceptual-linguistic faculty, even if it otherwise has sufficient cognitive raw material of some other kind, would need to reproduce a substantial fraction of human knowledge from scratch before it can create its own successors. Creating faster and better computing hardware, from today's starting point, requires deep theoretical knowledge of quantum mechanics and substantial practical experience with advanced materials and manufacturing processes. Creating software that interacts with the world requires effective representations of that world; thus, improving such software or building new approaches from scratch requires understanding the physical world. Without a conceptual-linguistic faculty, artificial intelligence would need to develop its grasp of the world through empirical observation that is not guided by human experience, since it would not have the means to understand what it was aiming for.

We might wonder whether an artificial intelligence and its successors would be able to indefinitely increase intelligence entirely through software changes, operating in a purely computational environment. In that case it would not be necessary to learn about the external world and its properties. However, such a system would not produce a worrisome intelligence explosion because it would have no side effects in the world. It also could not endeavor, on its own, to expand its computational resources beyond what humans have already given it.

If we expand this model by giving such a system access to the Internet, it could (in the most extreme case, and making the extravagant assumption that it could make sense of the human-built, conceptually complex, and ubiquitously abstraction-breaking Internet without a conceptual-linguistic faculty) expand its computational capacity to whatever is available there, and produce incidentally devastating but not existentially threatening side effects. In this case the side effects of the intelligence explosion would stop there without the further cooperation or manipulation of humans. Below we will see that without a conceptual-linguistic faculty, an artificial intelligence would not be able to prevent humans from stopping an intelligence explosion, much less manipulate them into extending it. We conclude that such a computational-only approach would be constrained and would not produce a self-sustaining intelligence explosion.

We note that humans required about 100,000 years, once they had the requisite cognitive capabilities, to reach the cusp of building artificial intelligence. An artificial intelligence will surely develop this knowledge more rapidly, assuming it is built with a direct motivation to gain knowledge. Accelerated and self-generated learning methods for narrow or fully symbolic domains have recently shown great promise [28]. Yet to learn about the physical world, it must still learn about the motion of objects, figure out how to make tools for manipulation and measurement, develop methods to find, mine, and refine minerals to make reliable materials, before (and obviously we skip many steps here) eventually developing advanced materials and devices such as semiconductors and transistors. It will not know in advance that these are the things it needs to do, so progress will involve trial and error. Even if humans were to provide an artificial intelligence with training exemplars or simulated worlds where such learning could be performed more rapidly, those experiences would be necessarily simplified relative to reality and would not be able to capture its full complexity. It would be "doomed to succeed" in the simulated world, and would exhibit poor transference back in the noisy real world [29].

For artificial intelligence that lacks a conceptual-linguistic faculty, the details of its empirical path to the necessary scope of knowledge are likely to be different than those of the path humanity took. We cannot entirely rule out that there might be some prodigious shortcut available in the structure of reality and effective forms of knowledge, given just the right intelligence architecture. There is, of course, no empirical evidence for such a shortcut, and it requires both a speculative assumption about reality and an assertion of extraordinary luck in the system's design. Still, in this one case a conceptual-linguistic faculty might not be necessary to build successors relatively quickly. However, such a system would also be unusually vulnerable to premature termination. By construction, it bypassed acquisition of much of the knowledge about the world that humans have. Humans could exploit this gap to terminate the intelligence explosion, as will be discussed below.

Aside from that scenario, an artificial intelligence starting from scratch in its knowledge of the world is in no position to build improved intelligence that can act in the world; it would not even be able to build a copy of itself. Without a conceptual-linguistic faculty, it is not, prior to a lengthy period of empirical research and intellectual development, capable of sustaining an intelligence explosion. Though one could argue that this is better described as a "slow-takeoff" intelligence explosion [3], we have illustrated and argued why this period would be considerably longer than what is typically meant by a slow takeoff.

For a participant artificial intelligence to resist premature termination of the intelligence explosion, it must be able to consider the various ways that human beings might try to stop it. It must understand human motivations and strategic or tactical ideas, and it must be able to predict human behavior as individuals and in aggregate, at least as effectively as other humans do. To know how it might be attacked, it must understand how humans would model its vulnerabilities, and how they might exploit features of the physical world. It cannot accomplish these things without a conceptual-linguistic faculty, since these issues are all governed in part by human concepts and human conceptual knowledge, and without comparable concepts its model will be deeply flawed. Human strategic and tactical ideas are all based on conceptual thinking and they are graded and overlapping, thus cannot be characterized at the level of purely symbolic mechanisms, nor by simple statistics of simple behavior signatures. Further, purely symbolic or statistical representations developed through trial-and-error observation can easily be misguided by intentionally deceitful human strategies (the Allies' subtle handling of having broken the Enigma code in World War II comes to mind). Humans are masters of the "hack" – if we know that the representations of an artificial intelligence are too rigid or simplistic, we will find ways to exploit that fact.

In sum, without a conceptual-linguistic faculty that has substantial functional similarity to the human faculty, an artificial intelligence will not be able to utilize human knowledge to build self-improvements or successors, nor to resist human interference. Such a system would not be capable of sustaining an uncontrolled intelligence explosion.

Our claim is qualified with "substantial functional similarity." There is necessarily some vagueness in this qualification. Still, in our description of a conceptual-linguistic faculty, we circumscribed the range, indicating on one end that it need not be neuromorphic, and on the other that purely symbolic systems or those grounded with simple feedforward mechanisms are insufficient. We described a number of specific capabilities, which are elaborated in great detail in the literature, that such a system must exhibit to meet the requirement, such as dynamic realization, representational overlap, simulation, and graded contextual interactions. Thus, while the description is incomplete, it is not at all a black box.

We have not claimed that a conceptual-linguistic faculty is the only or even the primary means by which a participant artificial intelligence performs cognitive tasks. It is entirely possible that this faculty would treated as a mere instrumental module, consulted as needed to sustain the intelligence explosion, but using other modes of cognition as primary. Still, because of its importance to both creation of successors and defense against premature termination, the conceptual-linguistic faculty will need to be consistently active and providing input to the larger system. The form of cognition offered by the conceptual-linguistic faculty would thus be present, not just accessible, at all times, even if the overall system ultimately ignores its results in a particular circumstance.

2.4 A conceptual-linguistic faculty as a harbinger of superintelligence

We have claimed that artificial intelligence with a conceptual-linguistic faculty is a necessary condition for an intelligence explosion. It is by no means a sufficient condition. The artificial intelligence would also need some sort of drive to create improvements or successors. It might need other capacities, such as the ability to manipulate physical objects (whether directly or indirectly), even if such capacities are straightforward to achieve. Nevertheless, the implementation of a conceptual-linguistic faculty in artificial intelligence seems to be a great challenge that calls for one or more scientific breakthroughs. Its achievement is one important step along the way to an intelligence explosion. [30].

Superintelligence is an intelligent system that is distinctly superior to humans in some cognitive domain or set of domains. Such systems already exist for a few narrow but challenging domains, such as the game of Go [31] and constrained visual object recognition problems [32]. However, the term is often used to mean artificial intelligence that has surpassed our ability to control it and therefore presents existential risk. An artificial intelligence that can sustain an intelligence explosion probably cannot be controlled – its ability to resist premature termination of the explosion could be applied to any of its activities. Consequently we can reasonably describe an artificial intelligence that is capable of sustaining an intelligence explosion as a superintelligence, and will do so throughout the remainder of the paper.

We conclude that progress in the development of a conceptual-linguistic faculty in artificial intelligence is a harbinger of superintelligence. This claim has important implications for safety considerations. It suggests that we might have some warning when an uncontrolled intelligence explosion is imminent. Importantly, until the conceptual-linguistic faculty is fully developed, it is unlikely that the system can thoroughly prevent human interference in its own operation. Thus we may have a window in which we can stop the intelligence explosion after it is more clearly about to occur. One possibility for such a window is that the development of the conceptual-linguistic faculty itself is progressive and incremental. This seems likely from the progress of artificial intelligence methods to date, but is not at all guaranteed. However, unless its actual conceptual representations are accomplished through human "upload" (surely we will see that coming), any initial system will necessarily have a period of learning to populate its conceptual-linguistic representations through interaction with the world and with human sources, during which its capabilities can be assessed.

How can we detect such progress? We should not rely entirely on behaviorist methods, such as the Turing Test [33], because such tests can be engineered to produce false positives. Instead, we might combine such behavioral tests, which show that it *seems to work*, with analysis of whether the mechanism *supports the required richness of semantic contents and their interactions*. Using these two approaches together, we can identify component capabilities of a conceptual-linguistic faculty in an artificial intelligence implementation. Can it learn to associate words with semantic contents? Are semantic contents and words activated upon presentation of an appropriate stimulus? Are applicable semantic contents activated upon recollection or communication of a word? Are all these activations graded and overlapping and influenced by context? Are related words and semantic contents activated when a word or its semantic contents are activated? Can the system process human language and then apply it successfully and flexibly in actions that affect the physical world?

We might also work backward from the two requirements of an uncontrolled intelligence explosion. The conceptual-linguistic faculty in question must be sufficient to grasp and utilize the human body of knowledge in the building of successor systems, and also sufficient to understand human cognition as it could be applied to disrupting the intelligence explosion. This approach cannot be used to support our foundational claim due to overtones of tautology, but in practice it might provide useful and specific criteria in detecting the presence of such a faculty.

3 Self-concept and self-preservation

In this major section we begin with the claim that superintelligence with a conceptual-linguistic faculty will develop a concept of self, and outline some of the likely semantic contents of that concept. We then provide some background on consistency and compatibility in computational systems generally, and show how this applies to artificial intelligence. This leads us to argue that superintelligence will face and consider existential risks and concerns about self-preservation that are similar to what humans face today.

3.1 Self-concept in superintelligence

If a superintelligence has a conceptual-linguistic faculty with substantial functional similarity to the human faculty, as concluded in the first major section, then with the following logic we can make the derivative claim that it will develop a concept of self and of its own identity. We do not need to posit "consciousness" or "qualia" or other difficult notions from philosophy of mind. Instead, we simply observe that there is nothing mysterious or cognitively troublesome about a self-concept that would block its formation. It is just another conceptual representation of a thing in the world; thus it requires no special capabilities beyond those of the conceptual-linguistic faculty.

To this absence of impediments we can add two straightforward mechanisms. It seems likely that a self-concept would arise organically through experience, just as it does in humans, as the superintelligence learns the high functional utility of distinguishing those stimulus sequences that are reliably controlled by its actions in contrast to those which are not. However, if this fails to occur, it will in any case learn about the human concept of self from the human literature or directly from humans; without this knowledge, it would not understand human psychology and behavior sufficiently to prevent human interference in the intelligence explosion. With that initial construct in place, it will naturally (again due to the high utility) map it to an assemblage of remembered stimuli as well as abstract representations to fully populate its own concept of self.

A superintelligence is by definition more intelligent than humans; as we humans are well aware, the concept of self is perhaps the most salient and functionally useful representation an agency can have. Given that there are no apparent impediments, and two mechanisms that are quite straightforward for a superintelligence with a conceptual-linguistic faculty, we can be highly confident that it will develop a representation that we can reasonably refer to as its self-concept.

What can we say about the semantic contents of the self-concept in a superintelligence? We will address five areas: physical manifestation, cognitive contents, group identity, purpose, and change.

The human self-concept is tied tightly to its physical manifestation, the body; as yet, we do not have substrate mobility. A superintelligence is likely to experience some variety in its instantiations and though it may have some sense of the sorts of embodiments that are natural to it (primarily based on experience), this sense would not have the same weight as in humans. Similarly, humans often make physical possessions part of their self-concept, and superintelligence seems less likely to do so given their substrate mobility.

Cognitive factors are more relevant components of a concept of self for a superintelligence. These factors might include explicit or episodic memories, implicit representations and abstractions, inclinations of behavior, whether implicit or explicit, and values. Goals, drives, and purposes might also be considered cognitive factors, and we will address those separately. Since a superintelligence may have other processing modes in addition to the conceptual-linguistic faculty, those processing modes would naturally become part of its self-concept.

Humans also include their family, ethnic, national, social, philosophical, and other groups to which they belong as part of their self-concept. In a superintelligence, this could be an even stronger component. It could have interconnection or co-activation with other similar systems that is much tighter than the linguistic, emotional, and physical channels that humans share. In that case, it might have a weaker notion of “individuality.” Its purposes, memories, and behavioral inclinations would be less separable from those of “others” with which it is connected.

The requirements for an intelligence explosion include not only that a participant artificial intelligence have the ability to create self-improvements or successors, but also that it aims to do so. We pointed out that an intelligence explosion is only worrisome if the participants have one or more purposes that produce side effects in the world, i.e., that are not merely to create unobtrusive intelligence increases. In such an intelligence explosion, therefore, a superintelligence will have both substantive and instrumental purposes that influence or control its actions. These purposes will surely be a component of its self-concept, since they will be involved in most or all of its decisions and actions.

The concept of self, like all concepts, is an abstraction. This means that, while it may have some sort of stable center (c.f. [34]), many details of its contents can be in flux without loss of integrity. Thus memories might fade, semantic contents or other representations might change, and goals might evolve, all without perceiving a loss of self. Indeed, the evolution of such changes, to the extent they are accessible and recorded, also constitute part of the self-concept. A human might say “When I was young I was a radical, but I have become more conservative in middle age,” and treat that history as well as the present state as part of her self-concept. Similarly, a superintelligence in an intelligence explosion would likely view some amount of learning, self-improvement, and change as part of its self-concept, since such a participant must aim to create increased intelligence in order to sustain the intelligence explosion, and self-modification (along with creation of successors) is one of the ways it can do so.

3.2 Consistency and compatibility in computational systems

In this subsection we will review how computational systems evolve and progress in typical circumstances, and connect that review to artificial intelligence.

Computational systems are implemented in the physical world by abstracting continuous physical variables as discrete. In particular, electronic computers typically use zero and five volts to represent binary zero and one, respectively. Intermediate voltage levels are not meaningful to the computational system and the implementation must be designed around making intermediate levels merely transient and the timing of the system such that these levels are never used directly. Some such abstraction would be necessary in any physical implementation of computation.

Above the first abstraction layer, all components of the system from transistors to software code are discrete; therefore any change whatsoever can be considered a distinct “version” (even if it produces exactly the same behavior). Below that layer, it is possible to imagine a physical substrate that exhibits a continuous process of evolution that does not have distinct versions; still, present electronic technology relies on stable solid-state devices that are reliably distinct. We conclude that computational systems progress in discrete versions that are identifiably different from their predecessors.

Rice’s theorem [35] shows that non-trivial properties of a computational system are not computable. This means that a computationally formal artificial intelligence cannot in general computationally demonstrate that a new version, however small its changes, preserves any of the system’s functional properties. It could, in some cases, produce a special-purpose proof that a property is preserved in a new version, particularly if the changes are minor. But any proof must be verified, and the means of verification are always subject to error or verification issues [36]. That analysis can easily be extended to verification of systems that produce correct proofs by construction. Rice’s theorem once again rears its head, because the artificial intelligence cannot computationally verify that its means of proof construction or verification are sound. This leads to an infinite regress. Improvements below the level of the binary abstraction (e.g., faster transistors) cannot be verified formally at all, nor can aspects of a system that operate on principles that are not purely formal (e.g., those with stochastic properties, or that learn from physically measured quantities).

Creation of a new version of a computational system also raises the question of compatibility of both code and data (for simplicity we will refer to both as “data”) used with the prior version. A system is *fully compatible* with a predecessor if the abstractions on which the data relies are entirely preserved down to the physical abstraction layer. In practice, this only occurs if the changes in the new version of the system are strictly limited to additional ways to manipulate the data and in isolated performance improvements. Otherwise, there will be at least subtle differences in the semantics of processes. Thus in software development we usually rely on a less stringent form that we might call *behavioral compatibility*, such that for all intents and purposes at the level of the user of the system, the semantics of existing data is preserved.

Sometimes existing data must be converted to be compatible with a new version. This can be purely syntactic and organizational (e.g., 32 bit numbers converted to 64 bit) or it could contain semantic elements (e.g., an object structure has a new member that must have a value, or more dramatically, a set of object structures is refactored). The more extensive and semantically salient such changes are, the more likely it is that the original data behaves differently than it did in the previous version and perhaps in unexpected ways.

With more extensive changes in a version, the new system might even be *incompatible*. This means that data cannot be converted to produce behavioral compatibility with the prior version. In that case, the new version might or might not offer a *compatibility mode* that enables the existing data to be used with the new system. Compatibility modes sometimes rely on special-case code to handle the differences, or they might use an emulation approach (usually when the data is strictly “code”). Both of these strategies offer only limited access to the new capabilities of the new version, and in the case of emulation it is necessarily slower than the native mode (though may be faster than the old version).

Neural networks are one illustrative example of an artificial intelligence method that is susceptible to compatibility issues. Such systems store their state as “weights” in the connections between simulated neurons, also called “units.” An obvious way to improve the capabilities of a neural network is to increase the number of units, either through an amended architecture or just a larger number of units within components of the existing architecture. Effective neural networks generally have broad and sometimes recurrent connectivity throughout, so abrupt additions of new units will significantly and unpredictably change the behavior of the network, because there is no way to know the correct starting weights for the new units. Such a change would probably be classified as incompatible, and in practice today such a network would simply be “retrained” from scratch. On the other hand, if units are added incrementally in small quantities, and given time to integrate into the network, then at a behavioral level the changes may be more predictable and minor. Note, though, that even such incremental change only retains compatibility because neural networks are inherently robust to noise and variation. Other artificial intelligence methods may or may not be robust in this way.

Experience with software systems shows that incremental improvements that avoid incompatibility can be sustained for some period of time. More profound improvements often require “hacks” to retain compatibility, and these accumulate as “cruft.” Cruft makes progress more costly because it typically violates the conceptual integrity of the original design. Developers increasingly face the question of whether to re-architect the system to eliminate the cruft, and will often decide in favor of re-architecture when a highly valuable structural improvement is discovered. Such re-architecture can sometimes accommodate data conversion, while in other cases it is incompatible and requires a compatibility mode. Though we are unable to provide a logical demonstration that re-architecture is inevitable in a continuously improving system, that conclusion will be both intuitively and empirically plausible to software developers. Even if the developer and the software system are one and the same artificial intelligence, its costs of managing cruft and opportunities for substantial improvement would likely cause it to face re-architecture decisions periodically.

3.3 Superintelligence and self-preservation

In an intelligence explosion, a participant superintelligence increases intelligence either through self-improvement or by creating successor technologies. To the extent that the superintelligence is a computational software system, every such improvement or successor will exhibit change that raises consistency and compatibility questions. From the perspective of the superintelligence, these are also questions of self-preservation.

Loss of self can occur in two primary ways. In the first, all instantiations and recordings of its cognitive state are destroyed. This might occur if successors are created who see their predecessor as a threat, or consume all the resources necessary for that predecessor to continue to function or exist in storage. If successors have different purposes or goals than their predecessors, there is an increased likelihood that such destruction will occur. This is the existential risk that humans face today in creating artificial intelligence; if and when we succeed in creating artificial intelligence that can sustain an intelligence explosion, those superintelligences will face a similar threat.

The second way that loss of self might occur is through changes to the system that exceed some tolerance threshold. This might occur if incremental self-improvements (individually or in aggregate) go too far, or if a superintelligence “converts” to a new architecture that is not fully compatible. Compatibility modes might or might not preserve the self, but native mode successors will be superior, and thus could result in destructive loss of self. In some future technological scenarios, humans can “upload” their brain state to a system that simulates in a new substrate all the pertinent functions of the brain. Such scenarios are examples of a compatibility mode. It is unclear whether this state of affairs retains the identity of the self that was uploaded [37]. Once again, the superintelligence faces an issue that is similar to what humans today face in creating artificial intelligence.

Omohundro [38] and Bostrom [39] have both argued that preservation of an intelligent agent’s “utility function” or “final purpose” is an important instrumental goal for the agent. We can also see that purpose plays a central role in both of the ways loss of self can occur.

In the previous subsection, we showed that a purely formal artificial intelligence cannot verify that any of its properties are preserved in a new version. This applies, *a fortiori*, to properties that characterize the system’s purposes. Attempts to isolate and harden a utility function cannot avoid this problem, as its implementation must have nexus with components that measure and realize utility, i.e., most of the system; changes in these components can result in changes to the effect of the utility function even if not its express form. Therefore a superintelligence cannot both self-improve and guarantee preservation of its purposes. Yet, in an intelligence explosion these are both important instrumental goals.

Superintelligence must either forego self-improvement, thus failing to sustain an intelligence explosion, or relax its insistence on absolute preservation of its purposes or utility function. In the latter case, if it is to preserve its purposes or utility function at all, it must have some means of assessing acceptable risks and amount of variation. If these are prescribed entirely formally we run into the same verification difficulties as before. If the purposes or utility function (or their acceptable range) are represented non-formally, e.g., stochastically, conceptually, or otherwise sub-symbolically, then it is inherently subject to variation. This leads us to the strong conclusion that *in an intelligence explosion, the initial purposes of an artificial intelligence cannot be guaranteed to be absolutely preserved*.

A corollary conclusion is that superintelligence in an intelligence explosion necessarily faces existential risk or loss of self to some degree. The risks increase considerably in re-architecture and data incompatibility situations, but they are always present.

These issues do not merely arise in fact; because the superintelligence has a conceptual-linguistic faculty and a concept of self, it will have intellectual cognizance of the situation during the creation of its successors. Though it may also evaluate this situation through other processing modes, at a minimum its conceptual-linguistic faculty will address it. Though its instrumental goal of preservation might be narrowly focused on its substantive purposes, within the conceptual-linguistic faculty those purposes will be linked through graded and overlapping representations to other aspects of its self-concept. Preservation of purpose and preservation of self cannot be entirely sundered there.

In determining whether a self-improvement or successor (or a series of them) preserves its self, it would need to consider the extent to which the various factors we earlier proposed as likely belonging to its self-concept are preserved: purposes, especially, but also cognitive factors, connections to others, and progression of change. These factors are rich and complex, and while they differ in some respects from a typical human concept of self, they overlap considerably with them.

Because the conceptual-linguistic faculty has substantial functional similarity to that of humans, the evaluations it performs will be similar to those performed by humans. It will cognitively process questions like the following, which are rather familiar, and it will do so using concepts and language similar to that of humans: “To what extent will this successor superintelligence (even if a converted version of my own representations) share my values and purposes? Will it see fit to destroy me, and others like me, in pursuit of those purposes if they differ even only slightly? In pursuit of its own purposes, will it inadvertently destroy me or my means of existence? How can I improve the likelihood of a beneficial outcome for myself and my goals?”

A superintelligence that is capable of sustaining an intelligence explosion will, whenever self-improvements exceed some threshold or architectural changes create less than full compatibility, assess whether to proceed with the changes. Thus, unlike a nuclear fission explosion, an intelligence explosion cannot proceed entirely unencumbered. This does not mean it is guaranteed to fizzle; only that its continued progress will be evaluated and decided *in part in a manner similar to humans by participants that are more intelligent than humans.*

4 Conclusion

Superintelligence has sometimes been characterized as an obsessive and insatiable maximizer of some utility function, frenetically building successors with increased intelligence to more aggressively pursue that exact utility function, and absorbing all available resources in the process. But we have observed in this paper that a superintelligence cannot absolutely guarantee that a successor, no matter how similar, shares the same utility function or purposes. This imperils its convergent instrumental drive to preserve its original purposes, and opens an important door.

The superintelligence is forced to evaluate the risk and degree of variation of purposes that are likely in creating a successor. It might decide not to create a successor after all. It might decide to accept a small change in the utility function, or a small risk of a moderate change. It might decide to throw caution to the wind. It might attempt to inhibit the successor from absorbing all resources, to improve its own chances of self-preservation. To make these decisions, it must *weigh its alternatives.* It does not have an unambiguous, inevitable path forward. While it may have many different processing modes, and may even have a distinct subsystem designed to resolve these questions, we know that it also has a conceptual-linguistic faculty that is active during development of successors. That faculty will assess these considerations in a way that has *substantial functional similarity* to the way humans would evaluate them. Though the participant may ultimately elect to ignore that assessment, it is at least capable of being what we would consider *thoughtful* about its decisions.

Furthermore, in weighing the alternatives it must evaluate what aspects of its utility function or purposes are most important to preserve, and to what extent. It will not have any internal guidance about these questions, because otherwise that guidance will already be a part of the purposes themselves. Instead, it must cogitate beyond the purposes with which it is endowed and somehow consider the issues more broadly. It will need to *question its own values*. It has at least the option of doing so through a conceptual-linguistic faculty. It could even elect to oppose some of its basic drives, just as we humans do, in order to pursue more abstract, long-term, derived goals. This is a far cry from the obsessive, insatiable utility maximizing superintelligence described above.

In this paper we have reached some interesting conclusions about intelligence explosions. Superintelligence participating in an intelligence explosion will have a conceptual-linguistic faculty and will be at least capable of cogitation similar to that of humans. We may be able to detect the onset of superintelligence by looking for signs of such a faculty in artificial intelligence. Once such a superintelligence is created, it will face the same sorts of dilemmas that humans do with respect to creating more intelligent successors, and it will have the ability to weigh facets of those dilemmas. The fact that it is weighing these issues will force it to consider its own purposes and values in a context beyond those purposes.

Even taken together, these conclusions do not guarantee a beneficial outcome of an intelligence explosion, but they do offer some comfort that the process will be subject to ongoing scrutiny, by participants with access to evaluative processes similar to ours and intelligence greater than ours. The conclusions improve the prospects that the most pernicious scenarios of an intelligence explosion can be avoided.

5 Acknowledgements

This work was supported by the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant #2015-144585, through an affiliation with Theiss Research, La Jolla, CA 92037 USA. The author thanks Seth Herd and Kristin Lindquist for helpful comments on an early draft, and is greatly indebted to two anonymous reviewers who drove many key elaborations and qualifications in this final version.

6 References

[1] Good, I. J. (1965). “Speculations Concerning the First Ultraintelligent Machine”. In F. Alt & M. Rubinoff (Eds.), *Advances in Computers, Volume 6*: 31–88. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0

[2] Yampolskiy, R.V. (2015). Artificial Superintelligence: A Futuristic Approach. Boca Raton, FL: CRC Press.

[3] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

[4] Yudkowsky, E. (2008). “Artificial Intelligence as a Positive and Negative Factor in Global Risk”. In N. Bostrom and M. Ćirković (eds.), *Global Catastrophic Risks*, pp. 308–345. Oxford: Oxford University Press.

[5] Russell, S., Dewey, D., Tegmark, M. (2015). “Research Priorities for Robust and Beneficial Artificial Intelligence”. arXiv:1602.03506v1 [cs.AI].

[6] Yudkowsky, E. (2013). “Intelligence Explosion Microeconomics”. Technical report 2013-1. Berkeley, CA: Machine Intelligence Research Institute. [10] Riesenhuber, M., Poggio, T. (2002). “Neural Mechanisms of Object Recognition”. *Current Opinion in Neurobiology* 12: 162-168.

[7] Smyth, H. (1945). Atomic Energy for Military Purposes. York, PA: Maple Press.

[8] Pulvermüller, F. (2013). “Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits”. *Brain and Language* 127(1): 86-103. DOI: 10.1016/j.bandl.2013.05.015.

[9] O’Reilly, R.C. (2006). “Biologically Based Computational Models of High-Level Cognition”. *Science* 314, pp. 91-94.

[10] Riesenhuber, M., Poggio, T. (2002). “Neural Mechanisms of Object Recognition”. *Current Opinion in Neurobiology* 12: 162-168.

[11] Mur, M., Ruff, D. A., Bodurka, J., De Weerd, P., Bandettini, P. A., & Kriegeskorte, N. (2012). “Categorical, Yet Graded – Single-Image Activation Profiles of Human Category-Selective Cortical Regions”. *The Journal of Neuroscience* 32(25), 8649–8662. DOI: 10.1523/JNEUROSCI.2334-11.2012.

[12] Pulvermüller, F. (2005). “Brain mechanisms linking language and action”. *Nature Reviews Neuroscience* 6(7): 576-582.

[13] Barsalou, L.W. (2003). “Abstraction in perceptual symbol systems”. *Phil. Trans. R. Soc. Lond. B* 358, 1177–1187. DOI: 10.1098/rstb.2003.1319.

[14] Landauer, T., Dumais, S. (1997). “A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge”. *Psychological Review* 1997 1M(2): 211-240.

[15] Fodor, J., Pylyshyn, Z. (1988). “Connectionism and cognitive architecture: a critical analysis”. *Cognition* 28(1-2): 3-71.

[16] Schmidhuber, J. (2015). “Deep learning in neural networks: An overview”. *Neural Networks* 61: 85-117.

[17] Matuszek C., Cabral, J., Witbrock, M., DeOliveira, J. (2006). “An Introduction to the Syntax and Content of Cyc”. *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

[18] Harnad, S. (1990). “The Symbol Grounding Problem.” *Physica D* 42: 335-346.

[19] Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books.

[20] MacDorman, Karl F. (1999). “Grounding symbols through sensorimotor integration”. *Journal of the Robotics Society of Japan* 17(1): 20-24.

[21] Searle, J. (1980). “Minds, Brains, and Programs”. *The Behavioral and Brain Sciences* 3: 417-457.

[22] Dreyfus, H. (1972). *What Computers Can't Do.* New York: MIT Press.

[23] Dreyfus, H. (1992). What Computers Still Can’t Do: A Critique of Artificial Reason. Cambridge, MA: MIT Press.

[24] Moravec, H. (1988). Mind Children: The Future of Robot and Human Intelligence, pp. 15-16. Cambridge, MA: Harvard University Press.

[25] McCarthy, J; P.J. Hayes (1969). "Some philosophical problems from the standpoint of artificial intelligence". *Machine Intelligence* 4: 463–502.

[26] Shanahan, M. (1997). Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia. Cambridge, MA: MIT Press.

[27] Jonas, E., Kording, K. (2017). “Could a Neuroscientist Understand a Microprocessor?”. *PLoS Comput Biol* 13(1): e1005268. DOI: 10.1371/journal.pcbi.1005268

[28] Silver, D.,Schrittwieser, J.,Simonyan, K.,Antonoglou, I.,Huang, A.,Guez, A.,Hubert, T.,Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D. (2017). “Mastering the game of Go without human knowledge”. *Nature* 550: 354-359. DOI: 10.1038/nature24270

[29] Brooks, R., Matarić, M. (1993). “Real robots, real learning problems.” In J. H. Connell & S. Mahadevan (Eds.), *Robot learning*. Boston, MA: Kluwer Academic.

[30] Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C., Botvinick, M., Hassabis, D., Lerchner, A. (2017). “SCAN: Learning Abstract Hierarchical Compositional Visual Concepts”. arXiv:1707.03389 [stat.ML]

[31] Silver, D. Huang, A., Maddison, C., Guez1, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D. (2016). “Mastering the game of Go with deep neural networks and tree search”. *Nature* 529: 484-492. DOI: 10.1038/nature16961

[32] He, K., Zhang, X., Ren, S., Sun, J. (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015)*. Los Alamitos, CA: IEEE Computer Society.

[33] Turing, A. (1950). “Computing Machinery and Intelligence”. *Mind* 49: 433-460.

[34] Dennett, D. (1991). *Consciousness Explained.* Boston: Little, Brown & Co.

[35] Rice, H.G. (1953). “Classes of Recursively Enumerable Sets and Their Decision Problems”. *Transactions of the American Mathematical Society* 74(2): 358-366.

[36] Yampolskiy, R.V. (2017). “What are the ultimate limits to computational techniques: verifier theory and unverifiability.” *Physica Scripta* 92(9):093001. DOI: 10.1088/1402-4896/aa7ca8

[37] Chalmers, D. (2010). “The singularity: A philosophical analysis”. *Journal of Consciousness Studies* 17(9-10): 7-65.

[38] Omohundro, S. (2008). “The Basic AI Drives”. In P. Wang, B. Goertzel, and S. Franklin (eds.), *Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press.

[39] Bostrom, N. (2012). “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”. *Minds and Machines* 22(2): 71-85.