Understanding and Controlling Artificial General Intelligent Systems

Jiří Wiedermann[1] and **Jan van Leeuwen**[2]

"The product of the human brain has escaped the control of human hands. This is the comedy of science."
Karel Čapek, 1920 [3]

Abstract. Artificial general intelligence (AGI) systems are advancing in all parts of our society. The potential of autonomous systems that surpass the capabilities of human intelligence has stirred debates everywhere. How should 'super-intelligent' AGI systems be viewed so they can be feasibly controlled? We approach this question based on the viewpoints of the epistemic philosophy of computation, which treats AGI systems as computational systems processing knowledge over some domain. Rather than considering their autonomous development based on 'self-improving software', as is customary in the literature about super-intelligence, we consider AGI systems as operating with 'self-improving epistemic theories' that automatically increase their understanding of the world around them. We outline a number of algorithmic principles by which the self-improving theories can be constructed. Then we discuss the problem of aligning the behavior of AGI systems with human values in order to make such systems safe. This issue arises concretely when one studies the social and ethical aspects of human-robot interaction in advanced AGI systems as they exist already today. No general solution to this problem is known. However, based on the principles of interactive proof systems, we design an architecture of AGI systems and an interactive scenario that will enable one to detect in their behavior deviations from the prescribed goals. The conclusions from our analysis of AGI systems temper the over-optimistic expectations and over-pessimistic fears of singularity believers, by grounding the ideas on super-intelligent AGI systems in more realistic foundations.

1 INTRODUCTION

The developments in robotics and AI and their extrapolations are pointing to the ever more credible scenario of highly advanced intelligent artifacts that outperform humans in all imaginable tasks. It is especially the role of philosophy to anticipate potential avenues for the further development of AI, envisage possible future problems, predict the related risks and look for ways how to harmonize the progress in AI with the interests of society. In their most advanced form, all these problems come concentrated in the general notion of superintelligence.

[1] Institute of Computer Science of Czech Academy of Sciences and Czech Institute of Informatics, Robotics and Cybernetics of Czech Technical University, Prague, Czech Republic email: jiri.wiedermann@cs.cas.cz
[2] Center for Philosophy of Computer Science, Utrecht University, the Netherlands email: J.vanLeeuwen1@uu.nl

Oxford philosopher and futurist Nick Bostrom defines superintelligence as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" [2]. Although it is not sure at all whether such a high level intellect can exist, and the more so, how superintelligence could be achieved, considerations about possible existential risks for mankind from constructing superintelligent systems possessing such abilities are abundant. They fill monographs, popular books, scientific papers, and general media and give scenarios for catastrophic sci-fi movies. Unfortunately, such considerations are rarely underpinned by more serious analyses based on findings from computability theory, computational complexity theory, the theory of AI and related sciences. Therefore the conclusions from such debates cannot usually be taken at face value.

Epistemic approach In this paper we approach the problem of understanding artificial general intelligence (AGI) from the viewpoint of the newly emerging philosophy of computation in which computational processes are understood as processes provably generating knowledge. This philosophy is presented in a number of recent papers by the authors, cf. [18], [19], [20], and [15]. Rather than analyzing HOW a computation is performed, the idea is to concentrate on WHAT is being computed, i.e., on what knowledge is generated in the course of computation (cf. [18] for the substantiation of this approach). So far, the framework has enabled us to deal formally with the problem of observer-relativity [19] and to tackle the problem of epistemic creativity [20]. A first formalization of the underlying theory with some basic results, is given in [15]. An overview of the approach is given in [20]. In the present paper we continue this line of research, now aimed at a deeper understanding of the self-improving mechanisms of AGI systems and the ways to control them.

Under the new view, knowledge generation becomes the hallmark of any computation and AGI and superintelligent systems both present but special cases of computational systems in which the ability to generate new knowledge is pushed to its maximum — such systems are able to generate knowledge over arbitrary epistemic domains corresponding to large parts of the sciences or the real world. This is in stark contrast with the practice of contemporary AI systems which are often specialized to particular, mostly quite restricted domains (cf. [12]). Also, each of them is usually developed in isolation from other systems. Such an approach does not easily support a unified view and hence, a unified theory of such systems.

Albeit quite general, our approach offers new insights in the nature of AGI systems. This is because the epistemic view of computation requires certain prerequisites that force a more detailed look at the underlying processes. The most important requirement is that behind any computation there must be a so-called *epistemic domain* over which computations are performed. The knowledge about a selected

part of this domain is described by means of an *epistemic theory* that can be more or less formal, or entirely informal. *Axioms* in this theory describe basic knowledge corresponding to the (representations of the) objects within the epistemic domain and of their basic properties. With the help of the *inference rules* the *derivations* within such a theory describe the ways how to construct objects corresponding to new pieces of knowledge. In a sense, new knowledge is generated by applying knowledge (in the form of inference rules) to old knowledge. The underlying computational processes are bound to the epistemic theory through the following condition: *whatever can be derived within the theory must be supported by the underlying computational processes (and vice versa).* If this condition is met, then what knowledge can and what knowledge cannot be generated over an epistemic domain and the "quality" of the generated knowledge (i.e., its compliance with the observations) depend solely on the underlying epistemic theory.

Self-improving theories The ability of a system to generate new knowledge in that system is thus catalyzed by the presence of mechanisms that allow changes in the underlying epistemic theory. Such changes can either extend the existing theory (by adding new pieces of truthful knowledge or new axioms), or "repair" it when a flaw is discovered in it — a discrepancy between the observed facts and the knowledge generated by the system. The ways how to construct such *self-improving theories* will be described in this paper.

The idea of self-improving theories stands in stark contrast with the generally accepted ideas in theoretical AI claiming that superintelligence should be achieved via so-called "recursively self-improving software" (cf. [2], [22]) that will help to overcome the level of human intelligence. Our modeling reveals that the idea of software self-improvement is not aiming at the core of the problem — in order to increase their intellectual potential, superintelligent systems have to continuously strive for increasing and repairing their knowledge, not software, related to their domain of interest.

Controlling safety The epistemic approach to computation opens interesting perspectives on controlling the *safety* of AGI systems w.r.t. the alignment of their actions with human values. Namely, we can imagine an AGI system that "observes" an other AGI system and checks whether its actions are safe. Unfortunately, no general solution of this problem is known, general in the sense that it would work for observing any AGI system. The respective control problem is widely considered in the "theory of superintelligence" [2] where it is called the *value loading problem*. In this respect Stuart Russell, in one of his interviews [11], posed an interesting, and deep question — namely, whether it is necessary to worry about undecidability for AI systems that rewrite themselves.

Nevertheless, we can offer at least a partial solution of the value loading problem. This solution is based on the existence of an observer that can detect a deviation in the behavior of an observed system from its original goals as described in the observee's underlying epistemic theory. Based on these ideas and the ideas of interactive proof theory we design an architecture of an AGI (or "superintelligent") system whose behavior can be controlled w.r.t. its alignment with its original values as given in its underlying epistemic theory. We also propose a scenario for interacting with such a system enabling its control and guidance toward the required results.

Perspective Before proceeding a word of caution is in place. Namely, from the context of our paper dealing with AGI and superintelligent systems one may get an impression that we are discussing matters at the frontiers of AI, to be current in a distant future when human level AI will be attained or even surpassed. But such an impression is misleading. That is to say, our findings have their say already now, in the current state-of-the art, e.g., in the context of robotics and related issues, like human-robot interaction and the social and ethical aspects of it. Obviously, each robot is controlled by some sort of AI system and, with the exception of simple fixed environments, the control system of robots must be flexible, able to adapt to new situations in a changing environment and under changing requirements of users. Hence, a control system for such robots must "tend" towards a full-fledged AGI system. Examples of such robots are robot companions, robots in personal care and health care, robots in search and rescue, etc.

Outline Interaction between humans and such robots is a necessary condition, and it is a natural requirement that all actions of robots should be justified w.r.t. the the current circumstance of the robots. This means that the actions and behavior of a robot must be rooted in some more- or less-formal "theory" capturing the epistemic domain in which a robot at hand operates. These ideas are at the heart of the epistemic approach to computation described in Section 2 of this paper. Next, a reasonable, although a highly non-trivial requirement to implement is that advanced robots must be able to adapt their behavior to the changing conditions in which they work. In our setting of epistemic computation, this prerequisite is translated into the requirement of self-improving epistemic theories that govern the behavior of the robots at hand. The principles of self-improving theories are described in Section 3. A *conditio sine qua non* when dealing with robots and in general, with AGI systems is, of course, safety. Any such system must not act against human interests. This is an intensively studied problem in social and ethical aspects of robotics. We discuss this problem in Section 4 where we indicate what obstacles lay on the road towards finding a reasonable solution of the problem of alignment a robot's actions with human values. Nevertheless, we will argue that, what can be done is checking that robot's behavior for its alignment with a given epistemic theory capturing robot's eligible behavior (Section 5).

All our results point towards the fact already noticed in the Čapek's quotation at the beginning of this paper — i.e., that time has come when we should see that the products of our brain — AI technologies — will not escape from our control. We have to change our approach to the technology risk from *reactive* to *proactive* (cf.[14]). Providing a balanced, well-founded and realistic view of the possibilities and limits of AI and robotics, of their benefits and threats, is indispensable for the positive social perception of the underlying technologies.

The structure of the paper is as follows. Section 2 is an introductory part giving brief preliminaries needed to understand the main ideas behind the epistemic approach to computation. In Section 3 we explain the idea of self-improving epistemic theories and sketch the mechanisms that take care of such abilities in AGI systems. The discussion concerning the existence of observers deciding the problem whether the behavior of an observed AGI system is aligned with the given set of values is given in Section 4. Section 5 deals with the possibilities of supervising AGI systems and devises a schema that, to a certain extent, solves the problem in a scenario asking for a continuous surveillance of a system's activities. Section 6 contains the conclusions.

2 AGI systems as knowledge generating systems

We assume that any AGI system is a computational system. We do not want to make any more specific assumptions concerning the nature of the underlying computations. In particular, we do not want to restrict our considerations to the currently known computational

systems. In order to make our approach independent of an underlying model of computation, we concentrate on the epistemic aspects of computations. Namely, we will not be primarily interested HOW a computation is performed, by what mechanism, by what computational model. Rather, we will be asking WHAT computations are doing. This is the approach taken in the so-called *epistemic theory of computations* coined by the authors in recent works, cf. [18], [19], [20], and [15].

The basic idea of this theory is the claim that computation is any process provably generating knowledge.That is, knowledge is the product of a specific process that, once it satisfies certain conditions, is termed as computation. These condition are: *compositionality* and *justifiability.* Compositionality means that a result of one computation can be used as an input to an other computation. Justifiability means that a computation must produce is results provably. The details are as follows.

Computation as knowledge generation A computation (i.e., the underlying process) works with knowledge belonging to some *knowledge domain* denoted as $\mathfrak{D}$. In many cases the knowledge domain is a representation of (a part of) the real world. $\mathfrak{D}$ can be given in a formal way — like in the case of mathematical or logical theories, or entirely informally — like in the case of knowledge domains described in a natural language capturing phenomena in the real word. Intermediate cases — a mix of formal and informal theories — are also acceptable. Theories may e.g. describe domains which are of interest in natural sciences like physics, biology, chemistry, etc. We even allow informal knowledge domains whose pieces of knowledge are pictures, paintings, music, poems, plans, maps, etc. In all cases, for each knowledge domain there must be so-called *rules of inference* which, when applied to pieces of knowledge, allow one to construct and to infer other pieces of knowledge.

The rules take the form of relations over pieces of knowledge. We require that each knowledge domain is closed w.r.t. the rules of inference (which means that by applying inference rules to a piece of knowledge one cannot get out of the underlying domain). A knowledge domain together with the corresponding inference rules forms an *epistemic theory.* The set of all axioms forms the *core* of $\mathfrak{D}$. In general, we do not require that an epistemic theory must be free of contradictions or that it must agree with all known facts (cf. [1] for similar ideas in the context of cognitive processing in organisms).

In order for a computation to generate knowledge there must be evidence (e.g., a proof) that explains that the computational process works as expected. The evidence must ascertain two facts: (i) that the generated knowledge can be derived within the underlying epistemic theory, and (ii) that the associated computational process generates it. The latter considerations are the key to the following, more formal definition (cf. [19]). (Do not forget that, although the notation used in the definition resembles the one used in formal theories, we will also be using it in the case of informal epistemic domains.)

Definition 1 Let T be a theory, let ω be a piece of knowledge serving as the input to a computation, and let $\kappa \in T$ be a piece of knowledge from T denoting the output of a computation. Let Π be a computational process. Then we say that process Π acting on input ω generates the piece of knowledge κ if and only if there is an explanation E such that the following conditions hold:

- $(T, \omega) \vdash \kappa$, i.e., κ is provable within T from ω, and
- E is a (causal) explanation that Π generates κ on input ω.

The device or mechanism realizing process Π is called a *computer.*

Note that our definition of computation is *machine independent*, since it holds for whatever process Π satisfying the conditions. It is also *algorithm free*, since we did not bother how process Π is realized. Last but not least, it is *representation independent* since the definition did not assume any particular representation of knowledge.

An important consequence of our definition is that knowledge is composable (since the processes generating knowledge are composable). An other important consequence is that a computation can be seen at various levels of abstractions. This is because process Π can again be seen as a process generating knowledge at a lower level of abstraction than implied by $(T, \omega) \vdash \kappa$. The former process can be seen as a computation working in a more fine-grained epistemic domain, consisting, e.g., of machine instructions. Of course, this view of levels of abstractions can be iterated, with computations on a given level of abstraction being implemented via computations at a lower level. For more details, cf. [15].

AGI systems Thanks to its generality our definition of computation holds not only for well formalized, so-called *theory-full knowledge domains,* but also for knowledge domains and inference rules that are hard to formalize. Such domains will be called *theory-less domains.* A typical example of a domain with informal inference rules is the domain describing the real world. Its objects, phenomena, actions in it and relations among the elements of this domain are described in a natural language. Knowledge about such domain can be described in the sentences of a natural language. The inference rules used in the natural language are the *rules of rational thinking and behavior.* These rules are based on the facts and reasoning that can be captured in a natural language. Typically, theory-less domains have extremely large knowledge bases (e.g., think of the contents of the Internet) and the inference chains (derivations) within such domains are relatively short.

The prime example of an intelligent system showing human-level intelligence is the brain, along with natural language. The brain supports inference processes in the informal theory formed by the natural language. In principle, instead of the brain one can think of any computer, even a computer not yet known or not yet existing. What we get will be an AGI system possessing human-level intelligence. It is the strength of our modeling of computations that it gives the means for working even with such poorly defined notions and comes with new understanding which has not been accessible through other approaches.

The view of AGI systems as systems provably generating knowledge brings an other bonus over the "classical" view of computations that sees superintelligence as actions of some super-algorithmic mechanism. Namely, Bostrom has reportedly argued that thought processes of superintelligent machines would be as alien to humans as human thought processes are to cockroaches [6]. Our approach reveals that this need not be the case: revealing the underlying epistemic theories and insisting on delivering proofs and explanations with each piece of derived knowledge would make the thoughts processes of AGI systems accessible to humans or at least to their current information processing technologies. We will make use of this fact in Section 5 when considering the possibilities of keeping the AGI systems under human control.

Note that the idea of controlling the operation of an AGI system via its proof checking works well in systems whose main task is "thinking", i.e., solving some purely intellectual problems. However, when an AGI system in question is a robot whose main purpose is to perform some prevailingly physical tasks requiring in addition to speech recognition and interaction a lot of vision and motor capabilities this approach is less suitable. This is because such tasks require

senso-motory and language skills that cannot, probably, be captured in a sufficiently formalized manageable theory that would drive the behavior of the robot at hand. In such cases the solution could probably be a hybrid approach delegating the realization of skills to specialized, pre-trained, "trusted" (deep) neural nets whose actions need not be checked each time when they are invoked. Their ability to perform the designated task follows from their construction after they were trained on large sets of data. The cooperation of these nets is then governed by some epistemic theory with checkable proofs (cf. [7] for a similar approach).

3 AGI systems with self-improving theories

Given an AGI system, the amount and quality of the knowledge it can generate depends on its underlying epistemic theory. A system endowed by the ability to *change* its underlying epistemic theory can thus influence its own "intelligence".

A system with a given epistemic theory can improve its ability to generate new knowledge in two basic ways:

(i) it can extend the core of its knowledge base by adding new knowledge to it, or
(ii) it can discover an "epistemic flaw" in its current epistemic theory and repair it by changing its flawed part or the entire theory.

The remaining option seems to be an extension of the repertoire of the AGI system's derivation rules by rules which could lead to shorter derivation chains. In our further analysis we will not concentrate on this option since centuries of practice in human reasoning have lead to a stable set of rules of reasoning, which indicates that this need not be reconsidered as a first option. There are many reasonable sets of inference rules known and all of them appear to be fit for use as far as their "derivation power" is concerned. Therefore, for knowledge derivation their change cannot have a dramatic effect measured in terms of "what can be derived". What a change of rules can influence is the efficiency of the inference process, but this is of secondary concern. However, it should not be overlooked that a system can change its *mode of reasoning*, i.e., change the type of underlying logic it uses. E.g., it can transit from purely deterministic reasoning to some probabilistic, fuzzy, modal, quantum, etc., way of reasoning. The use of non-standard logics is bound to specific epistemic domains (especially, in mathematical logic) and we will not consider it as a primary mean of amplifying the "power" of epistemic theories.

Changing theories We now consider the ways of changing epistemic theories in more detail.

(i) The first possibility of improving the ability of a system to generate new knowledge is the extension of the core of its epistemic theory. That is, adding of new knowledge to its core that cannot be inferred from the knowledge already existing in the knowledge base. Such new knowledge will play the role of new axioms. It can be obtained in two ways.

First, it can be delivered through sensors by which an AGI system interacts with its environment. Depending on the type of a sensor, the information provided by that sensor will be placed into the core as a new piece of knowledge. Note that what a sensor does is not computation, since a sensor realizes a mapping from the outside, "real world" into the knowledge domain $\mathfrak{D}$ (that is, a sensor does not implement a mapping from $\mathfrak{D}$ to $\mathfrak{D}$).

Second, new "ready made" knowledge, whose truthfulness is taken for granted, can enter an AGI system as an input from the outside. Such a piece of knowledge will become an axiom. Note that it shares the same characteristics as a piece of knowledge obtained through a sensor — it cannot be computed by the system at hand.

The extension of a theory in the above mentioned ways can be "automatized" as follows. An embodied AGI system, equipped with the necessary sensors and effectors, can by its own means propose and realize observations, measurements, experiments and in this way gain new information that can enter its knowledge base in form of axioms. Such a system can also ask an other systems (or humans) to assist it in getting the necessary information (cf. the scenario proposed at the end of Section 5). In any case, what happens is the injection of non-computable information (i.e., of information non-derivable in the given system) into the core of the epistemic base.

It is worth to observe that back in nineteen sixties M. Arbib [1] was already thinking along the lines of "how an intelligent organism may become more intelligent". To this end he invoked what he called *the Gödel speed-up theorem,* stating that if we add a new axiom to the existing axiomatization, not only are there truths which become theorems for the first time, but, also, that theorems which were already provable in the old system may have shorter proofs in the new system. In [1] he writes: *"If we make the highly artificial assumption that the contents of the memory of an organism corresponds to the axioms, and the theorems so far, of a formal system — with the ability to add axioms by induction — then Gödel's theorem reminds us that the virtue of adding new data to memory is not simply that that particular piece of information now becomes available to the organism, but also that the organism may compute other information much faster than it could otherwise, even though that information was denied to the organism before."*

(ii) The second possibility for improving the epistemic theory of a system is to discover an epistemic flaw in it and to repair it accordingly. E.g., a system may discover a discrepancy between its knowledge and some facts gained via observations and/or argumentation. This has been the case of the discrepancy between the geocentric and heliocentric theory. An AGI system can be designed so as to purposefully chase after such discrepancies, e.g., by actively verifying the "known truths" about the real world by its own means or in cooperation with humans (cf. the proposal of the interactive research scenario in Section 5). A system asked to find an explanation of some phenomenon can formulate a hypothesis — a missing link in a chain of explanations — and prove it by its own means. In order to do so a system must exhibit a high degree of epistemic creativity which, in this context, is seen as the ability to solve problems [21].

At this point a second historical digression is in place. Namely, the idea of *"falsificationalism",* i.e., of finding epistemic flaws in a theory serving as a refutation of that theory, has gained attention in philosophy since the work of K. Popper [10]. Popper envisioned the progress in science as a sequence of successive rejections of falsified scientific theories. In fact, our idea of "repairing" the flawed epistemic theories extends Popper's ideas from theory-full domains of scientific theories towards a more general setting of theory-less epistemic domains.

The combined effect of both theory extension and repairs has the consequence that the epistemic domain $\mathfrak{D}$ is enlarged and in this way the system keeps increasing its "understanding" of its environment. Could there be a better evidence for the intelligence growth of such an AGI system?

Returning back to Arbib's "highly artificial assumption", in the case of the AGI systems this assumption is becoming quite a natural one and in the form of self-improving theories leads to ... superintelligent artificial entities.

In addition to the previously described ideas of self-improving the-

ories, there is one more way of improving the efficiency of AGI systems without twiddling with the inference rules or the underlying theory. This is the idea of self-adjusting the often performed inferences w.r.t. the expected derivations. Such changes will not improve the worst cases, but can help in "average" cases. This can be implemented in many ways, e.g., by making the often derived intermediate results of inference a part of the knowledge base (in this way, later on they will be treated as axioms and there will be no need to repeat the respective derivations), by aligning the searches with the user preferences (as done by Google+), etc.

Some consequences There is a profound difference between our ideas of increasing the intelligence of AGIs and the generally accepted ideas of the *"recursive self-improvement"* of software. Intuitively, the difference between the ideas of "self-improving software" and "self-improving theories" seems to lie in the view of the underlying agencies: software is in principle uncontrolled and uncheckable, whereas theories are controlled and governed by checkable rules. But there is more to it: while the idea of self-improving epistemic theories is well founded in the epistemic approach to computations and can provably lead to qualitatively more powerful systems (through the increase of the amount of generated knowledge or the repair of incorrect theories), the idea of self-improving software leading to higher intelligence is quite vague and misguiding. Without a further proviso what parts of software are to be made more efficient and how, it can lead at best to quantitatively more efficient systems that would possibly be faster than the previous versions of the system. But such improvements will increase neither the knowledge capacity of the underlying system nor cure their knowledge flaws. Moreover, the often proclaimed idea of designing superintelligent systems that would have the ability of "getting better in getting better" cf. [22], [5], [2]) through software self-improvement has its limits derived from the limits of computations. The computational potential of physically realistic systems is upper-bounded by the Σ_2 class of the arithmetic hierarchy [16]. On the other hand, the idea of self-improving theories works as long as there are new aspects of the real or artificial worlds (like in mathematics) that are worth to be investigated. Our analysis clearly shows that what has to be primarily improved in order to gain greater intelligence, is not the software, but the epistemic data — the underlying knowledge base of the epistemic theories.

The latter considerations on self-improving theories also make more precise the ideas of so-called *Seed AI* (cf. [2]). Here, Seed AI should be seen as an AGI system which improves upon its own design, thus creating a smarter seed AI even more able to improve itself, and so on. Initially this program would likely have a minimal intelligence, but over the course of many iterations it would evolve to human-equivalent or even super-human intelligence. From our previous thoughts the following picture of Seed AI unfolds. Such a system must rest on the ability to extend and repair epistemic theories by some automated process, the ability to integrate the new improved theories in processes of learning, reasoning and action, shortening of its decision sequences and proofs, all of that supported by the interplay between a knowledge base and computation.

Our approach to superintelligence via self-improving theories allows the following proposition about the asymptotic convergence of the processes of well formalized epistemic theories:

Proposition 1 *In a stable well-formalized epistemic domain whose subsets are described by finite epistemic theories, in the long run all epistemic self-improving theories will converge towards theories generating equivalent knowledge.*

Proof (Sketch): Sooner or later, if there is a possibility to extend or repair the underlying epistemic theory, any AGI system based on such a theory will discover this possibility and improve its theory. In this way, all theories will keep disposing of their drawbacks. At the same time, of course, the theories keep extending their knowledge via refinement of the previously achieved knowledge. Assuming that theory improvement is confluent here, the AGI systems converge towards the equivalent epistemic theories under the stated conditions. □

It follows from the proposition that from a certain time on, sooner or later, any self-improving theory will only evolve as a so-called *normal science* (cf.[9]) in which no theory extensions or repairs will occur and the explicit knowledge generated by that theory will grow only through a refinement of the previously achieved knowledge. The proposition as it stands thus gives a certain support to J. Horgan's ideas concerning the end of science [8].

Proposition 1 may be compared with what Yampolskyi calls *"Recursive Self-Improvement Convergence Theory"* (RSI Convergence Theory) [22]. This theory hypothesizes that, regardless of the specifics behind the design of the Seed AI used to start an recursive self-improvement process, all such system, attempting to achieve superintelligence will converge to the same software architecture. While not specifying what is meant by the software architecture of a superintelligent system and lacking any theoretical framework, such a hypothesis must remain intuitive. Contrary to this, Proposition 1 is founded on the epistemic theory of computations and captures a different aspect of such an intuition — namely the one based on the epistemic view of AGI systems.

4 Observing AGI systems

The most tantalizing problem for AGI systems, and especially for superintelligent systems, is the issue of their *safety*. This is the question of how to construct AGI systems whose actions will be, and remain, aligned with human values (i.e., which would not harm human interests) while still allowing them enough freedom to change their goals in the course of their self-improvement.

There are two problems related to this question. First, there is no clarity on how to compose and represent a list of human values that AGI systems should preserve. As a first approximation the set of all laws, constitutions, human rights, treaties, religious books, etc., could be used. Obviously such a set might be be contradictory and incomplete. Determining any final form of it is bound to be full of political, philosophical and ethical issues.Therefore the implied theories implied will not be uniquely determined, subject to further (self-) improvement, and potentially diverge.

Unfortunately, even if we had an acceptable list of human values, there is a second problem: how to ensure that the behavior of the AGI systems will not deviate from the implied standard of well-behaved intelligence? No solution to these problems has been proposed until now [2]. Could it be possible to design a special AGI system $\mathcal{O}$ that would "observe" any other AGI system and raise an 'alarm' once it notices behavior that is going against the human values as agreed for this AGI?

Deciding safety In order to model this situation in our epistemic framework, let us think of two AGI systems working over the same knowledge domain, possibly with different theories. One is the system $\mathcal{A}$ — an AGI system to be observed, making use of theory $T_{\mathcal{A}}$, and the other is the system $\mathcal{O}$ observing $\mathcal{A}$ — with the goal of raising an alarm whenever $\mathcal{A}$ acts against human interests. We assume that $\mathcal{O}$ has an accepted list of human values, and that it has the ability to create hypotheses about $T_{\mathcal{A}}$.

In this context, "observing" means that $\mathcal{O}$ can observe the input/output behavior of $\mathcal{A}$, through appropriate sensors. It is conceivable that $\mathcal{O}$ can also access (a part of) the description, specification or even the 'code' of $\mathcal{A}$, for inspection. Does this mean that $\mathcal{O}$ can "decide" at all times whether $\mathcal{A}$ obeys the human values from its list? If there exist more observers of $\mathcal{A}$, with the same goals as $\mathcal{O}$ and equal 'lists' of human values, will their verdicts concerning the behavior of $\mathcal{A}$ always agree?

These problems are all very hard. To begin with, it is clearly impossible for $\mathcal{O}$ to decide whether $\mathcal{A}$ conforms to a given specification by just observing its input-output behavior for a limited time. Even code inspection may not be sufficient, as computability theory learns that it is generally impossible to decide non-trivial properties of a system from its code. This will be even more the case if $\mathcal{A}$ is allowed to self-improve, constantly renewing the theory that drives its actions. Thus, $\mathcal{O}$ seems forced to monitor $\mathcal{A}$ continually, and to keep track of any changes to $T_{\mathcal{A}}$ as they happen.

Even when it does so, the observer may have great difficulty to tell by observation whether $\mathcal{A}$ adheres to his list of human values. Adequate formalization seems required, but how to keep this compatible with changes in $\mathcal{A}$' behavior and signals? Without understanding the way $\mathcal{A}$ acquires new knowledge (and new behavior) $\mathcal{O}$ will never know if the AGI will continue to adhere to human values under all circumstances. Even code monitoring may not be sufficient, as behavioral properties may be impossible to decipher from code.

Due to the current lack of formalization and perhaps our poor insight we are not able to prove whether the properties which observers must "decide" are computationally undecidable. Nevertheless, we may counter this by attacking some of the assumptions on which the argumentation is based. Short of abandoning the whole idea that AGIs can be kept aligned with human values, the only alternative is to impose limitations and restrict their range of behavior.

5 Restraining the behavior of AGI systems

Although we are not able to prove the nonexistence of the general algorithm solving the value learning problem for any AGI that potentially can change its goals, we can prove the existence of an observer that can detect in the behavior of a given AGI system a deviation from its fixed goals, whatever they are.

In order to do so, the observer, having at its disposal the current theory governing the behavior of the observed system, simply checks whether the input/output behavior of the observed system is aligned with the behavior described by the theory at hand. Note the difference with the previous scenario — now the task of the observer is not to judge whether the behavior of the observed system is aligned with the list of acceptable behavior ("human values"). Rather, the observer has merely to check whether the observee does what is he expected to do according to the theory available to the observer.

Doing such a check, the observer need not replicate the observee's computations. Rather, for the observer it is enough to get from the observee the result of its computation together with the proof of the derivation of the result within its underlying theory. Checking such a proof is usually substantially easier than its discovery. Therefore, the observer can be a much simpler system than the observed system. In fact, the observer cannot generate any other knowledge than answers YES or NO concerning the correctness of the examined derivation. In particular, an observer cannot overwrite its theory.

Interactive proof systems We will make the previous thoughts more clear by using ideas from the theory of interactive proof systems. In computational complexity theory, an *interactive proof system* is an abstract machine that models computation by the exchange of messages between two parties. The parties, the *verifier* and the *prover,* interact by exchanging messages in order to ascertain whether a given string (fact) belongs to a language or not. The prover is all-powerful and possesses unlimited computational resources, but cannot be trusted, while the verifier has bounded computation power. Messages are sent between the verifier and prover until the verifier has an answer to the problem and has "convinced" itself that it is correct.

In our setting, the AGI system to be controlled will play the role of a prover while the checking observer will play the role of a verifier. The prover is all-powerful in the sense that it can generate whatever knowledge that can be derived within its epistemic theory T. Although the prover possesses unlimited computational resources it cannot be trusted — it can change its underlying theory (as explained in Section 3) and thus its derivations can deviate from those described by epistemic theory T that is available to the verifier. Thus, the "language" recognized by the verifier is the set of all proofs of statements in theory T. We assume that a bounded computational power of the verifier is enough for checking the proofs in theory T. In this way the verifier can detect whether the prover, in deriving its results, has used the original theory T (and thus, it did not change it), or not. This, of course, cannot prevent the prover to think about whatever it wants according to whatever theories it has developed. We do not care and we have no way of detecting it unless such ideas are presented to the verifier which will detect them as ideas not agreeing with the watched theory. Rather opposite is true — such thinking outside of the framework of the guarded theory may help the prover to increase its creativity and to find solutions that could possibly be transformed into the framework of the verifier's theory and presented to the verifier as proofs in the supervised theory.

Suppose now that the original AGI system $\mathcal{A}$ has been governed by a "fixed" theory T_0. This means that (so far) $\mathcal{A}$ has no mechanisms to modify T_0 and suppose further that we, humans, as creators of $\mathcal{A}$, after experiencing with $\mathcal{A}$ for some time come to a conclusion that this system is safe, aligned with human values. Now, let us allow changes of T_0. From this time on $\mathcal{A}$ is allowed to self-improve its underlying theory to get a different theory T_1. The verifier, having still theory T_0 to its disposal, can detect the deviation in behavior of $\mathcal{A}$ should the latter system use the derivations lying outside of T_0. This, of course, would not necessarily mean that $\mathcal{A}$ is no longer safe, nevertheless it could be the case.

Safe design The previous considerations together lead to the following design of a "safe" AGI system. The prover must be *boxed* (cf.[2]) so no information, except that serving for communication with the verifier, can escape from it. In order not to restrict its knowledge generating power giving rise to its creativity (cf. [20]), we assume that the prover has access to a fixed knowledge domain of some sort, or even to a copy of the entire Internet that is kept up-to-date. The prover starts its operation within the underlying epistemic theory T_0 that is considered to be safe by the operators of the system and is known to the verifier. The cooperation of both systems then proceeds in rounds. In each round, the verifier poses a "question" to the prover who has to find an answer and provide it to the verifier together with a proof that the verifier can check for correctness. Then the next round can start. This basic scenario can, of course, be made more detailed and can come in several variants. E.g., what to do if the prover is not able to find an answer in a reasonable time. Setting a time bound and allowing answers "I do not know" can solve this problem. We may also require that the prover will inform the verifier that it has discovered a new fact, or has "changed its mind" leading

to the change of its underlying epistemic theory. The operators of the system can then decide whether the new theory is safe or not. The prover can also propose to the operators to make an experiment and provide him with the results.

The scenario is guaranteed to be safe only when the following two assumptions are met.

(i) First, the quarantine of the AGI must be *perfectly tight*. For instance, if output of the prover could be accessible on the web, other AGI's could interpret the proofs as holding secret messages and thus conspire unnoticed to the verifier who believes that the box is tight. There are many other possibilities for scenarios of interactions between the prover, verifiers and operators of the entire system as well as many possibilities how information from the prover could leak out of the box. Thus, it appears that the true issue is whether a superintelligent system can always remain boxed in. E.g. unfriendly powers could release it. This may be unsolvable by our arguments but it does lead to the question how the architecture that is proposed can be enforced. A large number of possible scenarios and their strong- and weak points have been described in the Bostrom's book [2] in the section considering oracles and query-answering systems.

(ii) If the first assumption is fulfilled then the burden of verification rests on theory T_0. It must not allow any form of cheating from the side of prover. Especially, the verifier must not accept any false, incomplete "justifications" revealing only partial reasons for the solution proposed by the malicious prover not mentioning other reasons that might not be aligned with human values. A defense against such malicious attacks can only rest in a thorough inspection of all justifications provided by the prover and asking the prover for rendering the proofs that all aspects have been taken into account in the proposed solution.

Interaction When thinking about concrete interactions with superintelligent systems, it seems that any process of self-improvement should be guided and monitored with care. For example, it is probably not the best idea to want AGI's to solve complex problems in one shot — such as "cure cancer by discovering a process which eliminates cancerous cells from a human body without causing harm to the humans" (example taken from [13]), or "explain dark matter" [4].

A better idea seems to be to interact with the system, to devise a project for solving such a problem, a project counting on the cooperation with human teams of specialists. Such a proposal should consist of the introductory study on the status quo in the field at hand, identification of the current obstacles, indication of the promising ways to solve the problem, and their justification or explanation. After consulting this with humans, one could then interact and develop a methodology, a work plan stating what has to be done, what experiments and how and by whom they should be performed, stating hypothesis about what is to be expected, what will be the checkpoints, milestones, deliverables, risks, expenses, etc. Such an iterative incremental process will enable verification of the intermediate results, elimination of blind alleys, and above all — will keep AGI's under control, without trespassing human values (keeping us in the game as well as equal partners). The substantial novelty of our approach is the presence of the formal verifier which, together with the operators, can guard the safety of the system.

6 Conclusions

We have shown that the epistemic approach to computations enables a fresh look at AGI systems in general and at superintelligence in particular. This approach calls into question some generally proclaimed characteristics of superintelligence. First, it knocks down the idea that the thought processes of superintelligent systems will be so alien to human thinking that they become incomprehensible for people. The reason is that any system generating knowledge must have a proof that its inferences are correct within some epistemic theory, and this proof and the theory could be made accessible to humans. Second, the epistemic approach allows to identify the true source of self-improvement of superintelligent systems. It consists primarily in the self-improvement abilities of the underlying epistemic theories rather than in general and vague ideas about "recursively self-improving software". It is especially the epistemic data of which the quality and quantity must be raised, in order to improve the intelligence of AGI systems. Third, taking into account our inability to find an efficient solution of the alignment the AGI systems with human values, our approach to computation leads to new scenarios for constructing safe AGI systems. One such scenario, consisting in interactive monitoring the agreement of a system's actions with an acceptable epistemic theory and in intellectual partnership between humans and AGI systems, has been described in this paper. We believe that our conclusions moderate both the over-optimistic expectations of superintelligence fans and over-pessimistic concerns of superintelligence opponents, by grounding the ideas on singularity in more realistic foundations.

To conclude, the results of this investigation illustrate the strength of the epistemic approach to computation. In an elegant way, it allows a liberation from thinking in the frames of computational models, from the design of concrete algorithms underlying the actions of the AGI systems, and from the related representational issues. Under this approach, AGIs are seen as implementations of epistemic theories 'working' over epistemic domains that capture properties of the world. The resulting unifying framework allows one to study non-trivial general properties of AGI systems and leads to a deeper insight into the knowledge-producing mechanisms behind their operation. As far as we are aware, existing studies of AGI- and superintelligent systems (cf. [2]) do not offer such a general approach, with the same explanation potential.

ACKNOWLEDGEMENTS

The research of the first author was partially supported by ICS CAS fund RVO 67985807, Czech National Foundation Grant No. GA15-04960S and Research Programme of Czech Academy of Sciences Strategy AV21.

REFERENCES

[1] Arbib, M. A.: Automata Theory as an Abstract Boundary Condition for the Study of Information Processing in the Nervous System. In: Information Processing in The Nervous System. Proc. of a Symposium held at the State University of New York at Buffalo 21st 24th October, 1968, Editor: K. N. Leibovic, Springer 1969, pp. 3-19

[2] Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014

[3] Čapek, K.: An interview with the Czechoslovak author and playwright Karel Čapek (1890-1938) about his play R.U.R., London Saturday Review, 1920

[4] Deutsch, D.: Creative Blocks. In: AEON Magazine, 02 October 2012

[5] Chalmers, D.: The Singularity: A Philosophical Analysis. In: Journal of Consciousness Studies 17: 7-65, 2010

[6] Clever cogs: The potential impact of intelligent machines on human life. A review of the book [2]. The Economist, August 9th, 2014

[7] Garnelo, M., Arulkumaran, K., Shanahan, M., Towards Deep Symbolic Reinforcement Learning, arXiv preprint: 1609.05518, 2016

[8] Horgan, J.: The End of Science: Facing the Limits of Science in the Twilight of the Scientific Age. New York: Broadway Books, 1996

[9] Kuhn, T. S. The Structure of Scientific Revolutions. University of Chicago Press. 172 p., 1962

[10] Popper, K.: The Logic of Scientific Discovery, Basic Books, New York, NY, 1959

[11] Russell, S.: Concerns of an Artificial Intelligence Pioneer. In: Quanta Magazine, April 21, 2015

[12] Stone, P. et al.: Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report

[13] Soares, N.: The Value Learning Problem. MIRI technical report No. 2015-4, 2015

[14] Tegmark. M.: The Wisdom Race Is Heating Up. In: EDGE, a response to the question "2016: What do you consider the most interesting recent [scientific] news? What makes it important? https://edge.org/response-detail/26687

[15] van Leeuwen, J., Wiedermann, J.: Knowledge, representation and the dynamics of computation. To appear in: G. Dodig-Crnkovic, R. Giovagnoli (Eds): Representation and Reality: Humans, Animals and Machines, 2017, Berlin: Springer

[16] Wiedermann, J.: A Computability Argument Against Superintelligence. Cognitive Computation, September 2012, Volume 4, Issue 3, pp 236-245

[17] Wiedermann, J.: The creativity mechanisms in embodied agents: An explanatory model. In: 2013 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2013, pp. 41-45.

[18] Wiedermann, J. van Leeuwen, J.: Rethinking computation. In: *Proc. 6th AISB Symp. on Computing and Philosophy: The Scandal of Computation - What is Computation?*, AISB Convention 2013 (Exeter, UK), AISB, 2013, pp. 6-10

[19] Wiedermann, J., van Leeuwen, J.: Computation as knowledge generation, with application to the observer-relativity problem. In: *Proc. 7th AISB Symposium on Computing and Philosophy: Is Computation Observer-Relative?*, AISB Convention 2014 (Goldsmiths, University of London), AISB, 2014

[20] Wiedermann, J.,van Leeuwen, J.: What is Computation: An Epistemic Approach. (Invited talk). In Italiano, G.; Margaria-Steffen, T.; Pokorn, J.; Quisquater, J.J.; Wattenhofer, R. (ed.). SOFSEM 2015: Theory and Practice of Computer Science. LNCS 8939, Berlin : Springer, 2015, pp. 1-13

[21] Wiedermann, J., van Leeuwen, J: Towards a Computational Theory of Epistemic Creativity. In 41st Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB 2015). London: 2015, pp. 235-242

[22] Yampolskiy, R.V.: On the Limits of Recursively Self-Improving AGI. In: J. Bieger (Ed.): AGI 2015, LNAI 9205, Springer, 2015, pp. 394-403