The Singularity Institute

Artificial General Intelligence and the Human Mental Model

Roman V. Yampolskiy
University of Louisville

Joshua Fox
Singularity Institute Research Associate

Yampolskiy, Roman V., and Joshua Fox. Forthcoming. Artificial general intelligence and the human mental model. In *The singularity hypothesis: A scientific and philosophical assessment*, ed. Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. Berlin: Springer.

This version contains minor changes.

Abstract

When the first artificial general intelligences are built, they may improve themselves to far-above-human levels. Speculations about such future entities are already affected by anthropomorphic bias, which leads to erroneous analogies with human minds. In this chapter, we apply a goal-oriented understanding of intelligence to show that humanity occupies only a tiny portion of the design space of possible minds. This space is much larger than what we are familiar with from the human example; and the mental architectures and goals of future superintelligences need not have most of the properties of human minds. A new approach to cognitive science and philosophy of mind, one not centered on the human example, is needed to help us understand the challenges which we will face when a power greater than us emerges.

1. Introduction

People have always projected human mental features and values onto non-human phenomena like animals, rivers, and planets, and more recently onto newer targets, including robots, and even disembodied intelligent software. Today, speculation about future artificial general intelligences[1] (AGIs), including those with superhuman intelligence, is also affected by the assumption that various human mental properties will necessarily be reflected in these non-human entities. Nonetheless, it is essential to understand the possibilities for these superintelligences on their own terms, rather than as reflections of the human model. Though their origins in a human environment may give them some human mental properties, we cannot assume that any given property will be present. In this chapter, using an understanding of intelligence based on optimization power rather than human-like features, we will survey a number of particular anthropomorphisms and argue that these should be resisted.

2. Anthropomorphic Bias

Because our human minds intuitively define concepts through prototypical examples (Roach 1978), there is a tendency to over-generalize human properties to nonhuman intelligent systems: Animistic attribution of mental qualities to animals, inanimate objects, meteorological phenomena, and the like is common across human societies (Epley, Waytz, and Cacioppo 2007). We may use the term "anthropomorphic bias" for this tendency to model non-human entities as having human-like minds; this is an aspect of the Mind Projection Fallacy (Jaynes 2003). Excessive reliance on any model can be misleading whenever the analogy does not capture relevant aspects of the modeled space. This is true for anthropomorphism as well, since the range of human-like minds covers only a small part of the space of possible mind designs (Yudkowsky 2006; Salamon 2009).

In artificial intelligence research, the risk of anthropomorphic bias has been recognized from the beginning. Turing, in his seminal article, already understood that conditioning a test for "thinking" on a human model would exclude "something which ought to be described as thinking but which is very different from what a man does" (Turing 1950). More recently, Yudkowsky (2008, 2011) and Muehlhauser and Helm

1. The term "artificial general intelligence" here is used in the general sense of an agent, implemented by humans, which is capable of optimizing across a wide range of goals. "Strong AI" is a common synonym. "Artificial General Intelligence," capitalized, is also used as a term of art for a specific design paradigm which combines narrow AI techniques in an integrated engineered architecture; in contrast, for example, to one which is evolved or emulates the brain (Voss 2007). As discussed below, this more specific sense of AGI is also the primary focus of this article.

(forthcoming) have warned against anthropomorphizing tendencies in thinking about future superintelligences: those which surpass the human level of intelligence. To properly understand the possibilities that face us, we must consider the wide range of possible minds, including both their architecture and their goals. Expanding our model becomes all the more important when considering future AGIs whose power has reached superhuman levels.

3. Superintelligence

If we define intelligence on the human model, then intelligences will tautologically have many human properties. We instead use definitions in which intelligence is synonymous with optimization power, "an agent's ability to achieve goals in a wide range of environments" (Legg 2008). Legg uses a mathematical model in which an agent interacts with its environment through well-defined input channel, including observation and reward, as well as an output channel. He then defines a Universal Intelligence Measure (UIM) which sums the expectation of a reward function over an agent's future interactions with all possible environments. This definition is abstract and broad, encompassing all possible computable reward functions and environments.

Variations on this goal-based definition have been proposed. The Universal Intelligence Measure is so general that it does not capture the specific environments and goals likely to be encountered by a near-future AGI. To account for this, we can apply Goertzel's (2010) definition of "pragmatic efficient general intelligence," which resembles the Universal Intelligence Measure, but also takes into account the system's performance in given environments—which will likely be human-influenced—as well as the amount of resources it uses to achieve its goals.

There are cases where human-based definitions of intelligence are suitable, as when the purpose is to classify humans (Neisser et al. 1996). Likewise, human-based metrics may be applicable when the goal is to build AGIs intended specifically to emulate certain human characteristics. For example, Goertzel (2009) discusses practical intelligence in the social context: Goals and environments are assigned *a priori* probabilities according to the ease of communicating them to a given audience, which may well include humans.

Still, if the purpose is to consider the effects on us of future superintelligences or other non-human intelligences, definitions which better capture the relevant features should be used (Chalmers 2010). In the words of Dijkstra (1984), the question of whether a machine can think is "about as relevant as the question of whether Submarines Can Swim"—the properties that count are not internal details, but rather those which have effects that matter to us.

We use the term "mind" here simply as a synonym for an optimizing agent. Although the concept "mind" has no commonly-accepted definition beyond the human example, in the common intuition, humans and perhaps some other higher-order animals have a mind. In some usages of the term, introspective capacity, a localized implementation, or embodiment may be required. In our understanding, any optimization process, including a hypothetical artificially intelligent agent above a certain threshold, would constitute a mind.

Nonetheless, the intuitions for the concepts of "mind" and "intelligence" are bound up with many human properties, while our focus is simply on agents that can impact our human future. For our purposes, then, the terms "mind" and "intelligence" may simply be read "optimizing agent" and "optimization power."

Though our discussion considers intelligence-in-general, it focuses on superhuman intelligences, those "which can far surpass all the intellectual activities of any man however clever" (Good 1965). Superintelligences serve as the clearest illustration of our thesis that human properties are not necessary to intelligence, since they would be less affected by the constraints imposed by human-level intelligence. In contrast, the limited intelligence of near-human agents may well constrain them to have certain human-like properties. As research related to control and analysis of superintelligent systems gains momentum (Yampolskiy 2011a, 2011b, 2012a, 2012b, forthcoming; Yampolskiy and Fox, forthcoming) our thesis becomes essential for avoiding fundamental mistakes.

4. The Space of Possible Minds

Formalisms help broaden our intuitions about minds beyond our narrow experience, which knows no general intelligences but humans. In the approach mentioned earlier in association with the Universal Intelligence Measure, agents are modeled as functions which map, in repeated rounds of interaction, from an input-output history to an action (Hutter 2005; Legg 2008). As the inputs and outputs are modeled as strings from a finite alphabet, with an indefinite future horizon, there are, in principle, infinitely many such agent functions.

Using this model, there is a continuum across fundamentally different minds. For any computable agent, there are many other computable agents it cannot understand (learn to predict) at all (Legg 2006). There is thus, in principle, a class of agents who differ so strongly from the human that no human could understand them. Most agents represent trivial optimization power; our attention is focused on those which represent superhuman intelligence when implemented.

If we extend our model and allow the agent to change under the influence of the environment (Orseau and Ring 2011), we find another source of variety in alterations

in the mind itself, just as a given human behaves very differently under the influence of intoxicating substances, stress, pain, sleep, or food deprivation.

We are familiar with infrahuman intelligence in non-human animals. Animals can use senses, abilities such as navigation, and some forms of cognition, in goal-seeking (Griffin 1992). Non-human biological intelligences, including some different from those we are familiar with, could also evolve in environments outside our planet. Freitas (1979) describes an intelligence which might arise with a ganglionic rather than a chordate nervous system: such creatures, with small "brains" for each body segment (like most of earth's invertebrates), would have distributed, cooperative brains with distinct awareness for each body part. Likewise, animals with different weightings for their cognates of the three parts of the human brain—reptilian midbrain, limbic system, and neocortex—would have different distributions of mental features like aggression, emotion, and reason. Though these hypothetical biological intelligences fall into a narrow range of intelligence around the human level or below, they illustrate a range of possible architectures and motivational systems.

Classifications of kinds of minds which go much farther beyond the human example have been offered by Hall (2007) and Goertzel (2006). Hall classifies future AGIs, making the point that we should not expect AI systems to ever have closely humanlike distributions of ability, given that computers are already superhuman in some areas. So, despite its anthropocentric nature, his classification highlights the range of possibilities as well as the arbitrariness of the human intelligence as the point of reference. His classification encompasses hypohuman (infrahuman, less-than-human capacity), diahuman (human-level capacities in some areas, but still not a general intelligence), parahuman (similar but not identical to humans, as for example, augmented humans), allohuman (as capable as humans, but in different areas), epihuman (slightly beyond the human level), and hyperhuman (much more powerful than human). Goertzel classifies a broader range of minds, contrasting the human to possible non-human mental architectures, and describing AGI architectures which would implement many of these possibilities.

Singly-embodied minds control and receive input from a single spatially-constrained physical or simulated system; multiply- and flexibly-embodied minds, respectively, have a multiple or changing number of such embodiments. Non-embodied minds are those which are implemented in a physical substrate but do not control or receive input from a spatially-constrained body. Humans, of course, are singly-embodied.

Humans are not only embodied but also body-centered. The human brain is connected to and can be directly influenced by the remainder of the body, along with its immediate environment, so that the mind as a whole consists of patterns emergent between the physical system (the brain and body) and the environment. Non-embodied and non-body-centered minds are possible, and even within the narrower constraints of

embodiment, variations in the sensors and manipulators under control of a particular mind design present even more variety in mental capabilities.

Goertzel also explores possibilities for mind-to-mind linkage. Human minds work in near-isolation, connected mostly by the slow and lossy thought-serialization of language. But there are other possibilities. One is a mindplex, a set of collaborating units each of which is itself a mind. Human organizations and nations are mindplexes, albeit in imperfect form because of limitations in our communication; but a more tightly integrated mindplex would constitute a very different kind of general intelligence.

Within this variety of possible minds, superintelligence should not be considered a specialized variant of human level of intelligence. Rather, human-level intelligence should be considered an unstable equilibrium which can rapidly shift into superhuman ranges (Muehlhauser and Salamon, forthcoming). Humans find it difficult to improve their own brain power, but an AGI would find it much easier, since it would have capabilities such as adding more hardware or examining its own source code for possible optimizations. Moreover, most AGIs would want to self-improve to the highest possible level of intelligence, as this has value in achieving most goals (Omohundro 2008).

Humans are the first general intelligence on earth. We have been in existence for a short time in evolutionary terms and represent a lower bound on the intelligence able to build a civilization. The upper limit on raw processing power for the entire universe through its history, as imposed by the laws of physics, 10^{120} operations over 10^{90} bits; a one-liter, one-kilogram computer has the upper limit of 10^{50} operations per second (Lloyd 2000, 2002). This theoretical maximum is almost certainly too generous—it assumes an exploding computer. But even with tighter constraints, such as speed of electrical and optical signals in feasible technology, or the Landauer (1961) limits on the minimal energy required for any irreversible computation, upper bounds remain far above the human level, estimated at 10^{11} operations per second (Moravec 1998). Even though functional ability requires more than just raw power, the gap between the human level and the highest degree of optimization power possible leaves open a wide range, encompassing a vast range of possible superhuman intelligence levels (Sotala 2010).

5. Architectural Properties of the Human Mind

The tight entwinement of functionality and goals in the human brain is a contingent fact which depends on our evolutionary history. But in general, a single architecture may serve various goals, while multiple architectures may be capable of serving a given goal system. Thus, mental architecture and goal systems must be examined separately.

The human mental architecture is quite uniform, the so-called "psychic unity of mankind." This is a result of humanity's origin in evolution through sexual reproduc-

tion, which works only when genomes remain similar across the species. This results in homogeneity in human minds, both in hardware—the brain—and software—the functional mind design (Tooby and Cosmides 1992). All human minds share specialized features and behaviors, including myths, grammar, ethnocentrism, play, and empathy, and many others (Brown 1991, 2004). Other animals, which share a biological substrate and the goals of reproductive fitness with humans, also share certain human mental features, as for example, specialized abilities to track degrees of genetic relatedness. But non-biological optimizers, which are not faced by these constraints, need not have the same motivations and accompanying mental techniques.

For humans, the perceptions of space and time, and the ability to act on their environment, are centered on a body. Embodiment-based cognition is so essential to human minds that it extends even to aspects of cognition which do not directly depend on embodiment (Lakoff 1987). For example, one "wades through" a difficult book. An AGI must likewise be implemented in some physical form (which must be protected if the agent is to continue working towards its goals). It also must interact with its environment in space and time if its goals are based on the state of the environment. Thus, perceptions and action are also essential to a superintelligence. Body-centeredness is not necessary, however, since a computer substrate allows the distribution and relocation of mental capacities, perceptions and motor control.

Human minds are characterized by some weaknesses. Even ordinary computers surpass us in symbol processing and logical inference. Humans, for example, are typically unable to trace nested (non-tail) syntactic recursion to more than about two levels (Reich 1972; Karlsson 2010), though computers can do this with ease; a superintelligence could easily adopt such a computational module.

Minds like ours, which work with very limited computational resources, have to rely on heuristic simplifications to arrive at satisficing solutions. These heuristics create biases which constitute imperfections in human rationality (Kahneman, Slovic, and Tversky 1982; Gigerenzer and Selten 2001). For example, humans are afflicted with the endowment effect, in which possessions which one currently owns are overvalued, so that investors often avoid selling assets, where that would maximize utility. A near-human or only slightly super-human artificial intelligence might also find similar heuristics to be necessary. But human biases are not necessary to intelligence. Indeed, even without superintelligence, the narrow AI financial systems of today can ignore the endowment effect in making buy/sell decisions. A superintelligence with adequate computational power and memory would not need to adopt these heuristics at all.

A superintelligence with origins in a designed AGI, rather than in evolution, will lack the weaknesses of a biological substrate. While the human brain is constrained by its evolutionary origins, engineers have available to them a far wider range of designs than

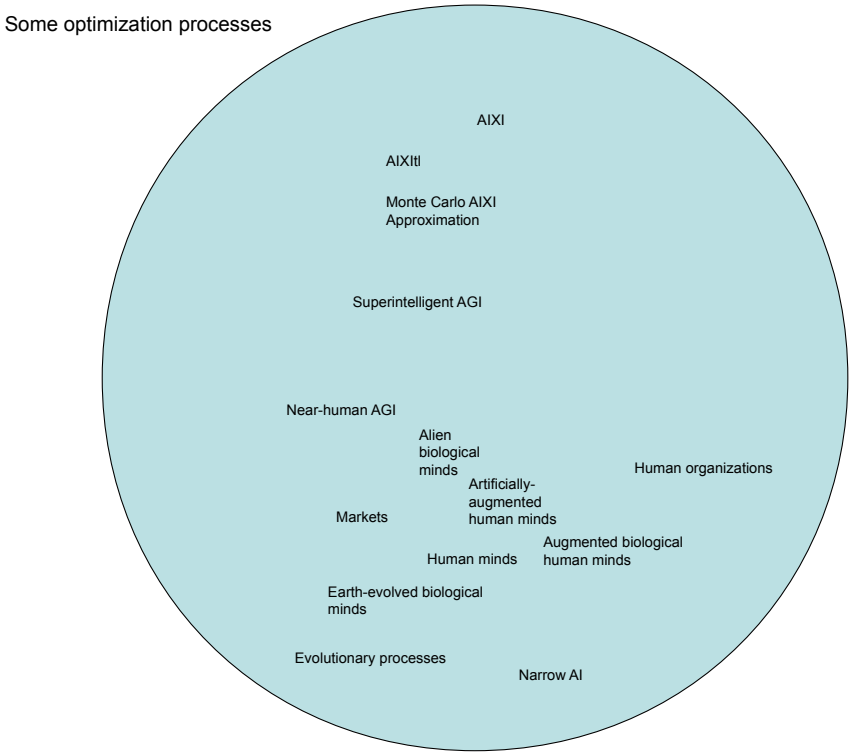

Figure 1: This figure, based on Yudkowsky (2006), presents the optimization processes surveyed here, including existing optimizers, hypothetical ones, and formalisms. (The figure is not to scale and shows only an unrepresentative sample of the possibilities.) It is intended to illustrate that human mind design constitutes a tiny part of a vast space of possible minds, most of which have deeply non-human-like goals and architectures.

evolution did in sculpting us (Bostrom and Sandberg 2009). Engineers have the benefit of memory, foresight, cooperation, and communication; they have the ability to make leaps in design space, and trial-and-error cycles on a short time-scale (Yudkowsky 2003). A computer substrate also provides advantages over human brains, such as modularity, better serial computation, and transparent internal operation.

Many weak-AI implementation projects have incorporated human properties such as embodiment, emotions, and social capacities (Brooks 1999; Duffy 2003). Some projects go so far as to intentionally copy limitations observed in human psychology, in order to avoid wasting effort on potentially unfeasible tasks, while still achieving human-comparable performance (Strannegård 2005). However, implementations of isolated mental features give us no reason to assume that a full AGI would necessarily have a wide range of human-like properties.

Already today, many forms of narrow AI and proposed designs for strong AI have non-anthropomorphic, computer-based architectures. For example, the design paradigm called Artificial General Intelligence adopts narrow-AI components but takes only a

broad inspiration from the human mind (e.g., Goertzel et al. 2009). The variability of the architecture will be all the greater after self-improvement, since an AGI need not keep its current architecture as it self-improves. It will create a new design, even a radically non-anthropomorphic one, if an entity with this new design will be better able to achieve its goals.

There is a category of greater-than-human minds which *would* have human-like mental properties: those derived from humans. These include brains augmented with nootropic drugs, genetic engineering, or brain-computer interfaces (Vidal 1973; Graimann, Allison, and Pfurtscheller 2010), uploads of specific persons (Hanson 1994), and whole brain emulations (Sandberg and Bostrom 2008). On functionalist principles, these are all rooted in and not essentially different in their origins from ordinary biological minds, and inherit their properties, even if subsequently they use their greater-than-human intelligence to bootstrap to much higher levels. Thus, these human-origin superintelligences are likely to present examples of human properties. Still, fundamental differences from human architecture can be expected. Uploads, human/machine cyborgs, and whole brain emulations, with their non-biological computing substrate gain advantages over biological brains in areas such as self-improvement, communication, and others (Sotala, forthcoming). As they improve to superintelligence, the human-origin minds could leave behind human mental limitations, and reimplement themselves in an even better architecture if they so choose.

6. Human Goals

The human goal-system, which includes survival, social status, and morality, along with many others, is a mix of adaptations to conditions in the human ancestral environment (Tooby and Cosmides 1992; J. D. Greene 2002). In contrast, an AGI, and in particular a superintelligence, can have arbitrary goals, whether these are defined by its designers or develop in a random or chaotic process.

Human terminal values arose from their instrumental value in achieving evolution's implicit goal of reproductive success for the genes. For AGIs as well, such human-like preferences would have instrumental value for the achievement of many goals. For example, most agents, including superintelligences, would be motivated, as humans are, to protect themselves, to acquire resources, and to use them efficiently.

But there are also instrumental values which could be of far more use to machine intelligences than to humans. Humans can take nootropic drugs to enhance their minds; they can avoid addictive or psychoactive drugs to avoid distorting their utility function. But an agent which can more fully examine and improve its own design, implementation,

and embodiment will find much more value in self-improvement, preservation of utility functions, and prevention of counterfeit utility (Omohundro 2008).

An intelligence far more powerful than humans would have no need for the values of exchange, cooperation, and altruism in interaction with humans, unless these were built in as terminal values. A superintelligence would not need any benefits that humans could offer in exchange for its good behavior; it could evade monitoring and resist punishment. Since humans cannot meaningfully help or harm the superintelligence, there would be little value in cooperation, verifiably trustworthy dispositions, or benevolence (Fox and Shulman 2010), nor in money (control of resources on a human scale), social power, or even malevolence (disempowerment or elimination of rivals as a goal).

In an environment with peers, a superintelligence would have incentive to cooperate or compete instrumentally (Hall 2007). Yet this is only true where superintelligences of roughly equal ability exist. Across the wide continuum of possible levels of intelligence, agents which are not of the same species—the same software/hardware specification—are more likely to be mismatched than to be equals. A Darwinian scenario, in which a population of superintelligences cooperates and competes, could produce rough equals, perhaps distributed across niches. But if a superintelligence self-improves fast enough, it will be aware of the evolutionary threat and suppress the rise of other intelligences (Bostrom 2006).

Human goals are mostly self-centered, with altruism a debatable exception (Batson 2010). In contrast, there is no *a priori* reason for AGIs to treat their own control of resources or their own continued existence as terminal values, athough they may be useful as instrumental values. Since future AGIs will have goals designed to serve human preferences, they may well possess a quite inhuman altruism.

If today's plans for AGIs are any guide, the first ones are likely to be assigned simple goal-systems, as contrasted to the human multiplicity of mutually inconsistent, changeable goals, with intertwined instrumental and terminal values. (However, complex goal systems are possible in AGIs, particularly if their creators are specifically trying to copy the human goal system.)

AGI goals originating in human needs (for example, maximizing wealth or winning a war for its makers), are no guarantee of human-like behavior, even if these goals are well-defined. A drive to maximize these goals, even at the expense of all other values important to humans, would result in deeply alien behavior (Bostrom 2003; Yudkowsky 2011).

Humans sometimes change their values, as in a process of Kantian Reflection, in which a person decides that moral reciprocity is not merely a means to an end, but also an end in itself. However, any sufficiently powerful superintelligence would not change its values, since doing so impairs the chances of achievement of the current values, and

so represents a limitation to optimization power. Thus, a very powerful optimizer would strive to prevent such human-like preference evolution (Omohundro 2008).[2]

Superintelligences originating in humans, such as augmented brains, cyborgs, uploads, and whole brain emulations, would start with human values. As they gain in power, they would lose the social constraints which form an important motivation for human behavior: Power corrupts, and power far beyond what any tyrant has known to date may corrupt a human-like mind so that its motivational system becomes very different from that of today's humans. As human-based minds self-improve, they are likely to seek to protect their goal systems, like other powerful optimizers; this could produce an example of a superintelligence with human-like goals. Even these superintelligences, however, may ultimately evolve goals which differ from those of humans. They would start with human-like changeability in their goal system. The changeability could in itself be treated as a meta-value, resulting in different object-level values.

7. Examples of Superintelligences

We know of no superintelligences today, nor any other intelligences with the generality and flexibility of the human mind. But examples of powerful optimization processes with non-human goals and architectures are available. Some are superior to humans in certain areas of optimization.

Evolutionary processes are flexible and powerful enough to create life forms adapted for a wide variety of changing environments by optimizing for reproductive success, far outdoing the accomplishments of human engineers in this area. (It should be noted, however, that these processes have had much more time to work than all human engineers combined). Evolutionary processes share with humans only the ability to optimize; they lack all other properties associated with humans. They are unembodied, impersonal, unconscious, and non-teleological, lacking any modeling capacities. Even though the human mind evolved to serve evolutionary goals of reproductive success, humans do not share the goals of the evolutionary processes which created them (Cosmides and Tooby 1992; Yudkowsky 2003).

2. Change of goals is possible in a superintelligence where a stable metagoal is the true motivator. For example, discovery and refinement of goals is part of Coherent Extrapolated Volition, a goal system for a self-improving AGI. It is designed, to ultimately converge on the terminal value of helping humans achieve their goal system as extrapolated towards reflective equilibrium (Yudkowsky 2004; Tarleton 2010; Dewey 2011). Nonetheless, CEV does not violate the principle that a sufficiently powerful optimizer would lack human-like variability in its goals, since its meta-level values towards goal definition in themselves constitute a stable top-level goal system.

Markets are another type of powerful optimizer. Though externalities and other market failures render them far from optimal, they outdo centralized planning (i.e., a small group of human minds) at their implicit goal, maximizing for the net benefit of producers and consumers. Though based on the interactions of individual humans, each working towards their own goals, markets as a whole lack all properties of the human mind. Markets are embodied in the humans who participate in them, but optimize distinct values from any individual human. Like evolutionary processes, markets are impersonal, unconscious, and non-teleological; and lack internal models of the optimized domain.

Markets present a valuable example of other-directed goals: They optimize functions which are aligned with and derived from the values of other intelligences, namely humans. Such other-directedness is rare in humans and other biological intelligences. In contrast, in artificial agents there is no bar to pure altruism; in fact, since they would be created to serve their designers' goals, other-directed values are the default.

These examples are useful, but limited. None is a true superintelligence. Only humans today have flexible, general, intelligence, leaving theoretical models of superintelligence such as AIXI as a useful tool in considering the full range of possibilities. AIXI (Hutter 2005) is an abstract and non-anthropomorphic formalism for general and flexible superintelligence. It combines Solomonoff induction (Li and Vitányi 1993, 282–290) and expectimax calculations to optimize for any computable reward function. It is provably superior at doing so, within a constant factor, than any other intelligence (Hutter 2005).

There are limitations on the usefulness of AIXI as an example. As it is incomputable, it must be treated as a model for intelligence, not as a design for an AGI.

AIXI, and Legg's Universal Intelligence Measure which AIXI optimizes, is incapable of taking the agent itself into account. AIXI does not "model itself" to figure out what actions it will take in the future; implicit in its definition is the assumption that it will continue, up until its horizon, to choose actions that maximize expected future value. AIXI's definition assumes that the maximizing action will always be chosen, despite the fact that the agent's implementation was predictably destroyed. This is not accurate for real-world implementations which may malfunction, be destroyed, self-modify, etc. (Daniel Dewey, pers. comm., August 22, 2011; see also Dewey 2011). AIXI's optimization is for external rewards only, with no term for the state of the agent itself, it does not apply to systems that have preferences about more than the reward, for example, preferences concerning the world as such, or preferences about their own state; nor does it apply to mortal systems (Orseau and Ring 2011). Nonetheless, AIXI does a good job of representing the best possible optimizer in the sense of finding ever closer approximations to the global maxima in a large search space of achievable world-states.

Taken as an abstract model, AIXI's complete and compact specification serves to show that in the limiting case, almost any property in an intelligence, beyond optimization power itself, is unnecessary.

AIXI is quite inhuman. It is completely universal, maximally intelligent under a universal probability distribution—i.e., where the environment is not prespecified. It thus lacks the inductive bias favored by humans. It lacks human qualities of embodiment: It has no physical existence beyond the input and output channels. Also, unlike humans, this formalism has no built-in values; it works to maximize an external reward function which it must learn from observation.

Variants of AIXI bring this model, with its compact specification and freedom from built-in inductive bias, into the realm of computability and even implementation. AIXItl (Hutter 2005) is computable, though intractable, and is provably superior within a constant factor to any other intelligence with given time and length limits. A tractable approximation, Monte Carlo AIXI, has been implemented and tested on several basic problems (Veness et al. 2011).

8. A Copernican Revolution in Cognitive Science

We have explored a variety of human mental properties, including single embodiment, body-centeredness, certain strengths and weaknesses, and a specific complex set of goals. Superintelligences need have none of these features. Though some instrumental goals will be valuable for most intelligent agents, only the definitional property of much-higher-than-human optimization power will necessarily be present in a superintelligence. Humans are the only good example of general intelligence which we know—but not the only one possible, particularly when the constraints of our design are thrown aside in a superintelligence.

Since the Copernican revolution, science has repeatedly shown that humanity is not central. Our planet is not the center of the universe; *homo sapiens* is just another animal species; we, like other life-forms, are composed of nothing but ordinary matter. Recently, multiverse theories have suggested that everything we observe is a tiny part of a much larger ensemble (Tegmark 2004; B. Greene 2011).

This decentralizing trend has not yet reached the philosophy of the human mind. Much of today's scholarship takes the universality of the properties of the human mind as granted, and fails to consider in depth the full range of possible architectures and values for other general optimizers, including optimizers much more powerful than humans. It is time for psychology and the philosophy of mind to embrace universal cognitive diversity. Even in today's era, in which only a single design for general intelligence exists,

this broadening will enrich our analytic tools for understanding mental architecture, decision processes, goals, and morality.

A Copernican revolution for the mind can extend our view outwards, but also improve our insight into ourselves. The shift away from geocentric cosmology improved our understanding of the Earth, and an evolutionary analysis of our species' rise helped us understand the design of humans. So too, an examination of other possible minds, and in particular superintelligent minds, can help us reach philosophical and psychological conclusions about humans as well. Already, infrahuman AGIs have provided paradigms for philosophy of mind (e.g., Newell and Simon 1972); AI-related research such as Bayesian network theory (Tenenbaum, Griffiths, and Kemp 2006) has contributed to neuroscience. Though at the current stage only thought experiments are possible, theories about possible superintelligences can shed more light on the human condition.

Today's astronomers know that Earth is still central in one sense: We live on it; our observations are made from it or near it; our tentative explorations of space began on it; its fate is tied up with our own. So too, when we humans begin exploring mind-space with the creation of AGIs, the human mind will remain of central importance. The first near-human-level AGIs may be partially modeled on our mind's architecture and will have goals chosen to serve us. But just as astronomers came to learn that the universe has unimaginably large voids, stars much greater than our sun, and astronomical bodies stranger than anything previously known, so too we will soon encounter new intelligences much more powerful than us and very different from us in mental architecture and goals.

There are two meanings to Copernicanism. One is "we are not central," and the other is "we are ordinary; what we see is common." This second meaning, too, should influence our thinking on intelligence. Although the human mind's special status as the only true general intelligence remains a reality for now, in principle other general intelligences can exist. Once other human-level intelligences, and then superintelligences, are created, our theory of mind will have to expand to include them; we should start now, arming ourselves with an understanding which may enable us to design them to meet our needs. Defining the initial AGIs' goals in accordance with human values, and guaranteeing the preservation of the goals under recursive self-improvement, will be essential if our human values are to be preserved (Yudkowsky 2008; Anissimov 2011).

Acknowledgments

Thanks to Carl Shulman, Anna Salamon, Brian Rabkin, Luke Muehlhauser, and Daniel Dewey for their valuable comments.

References

Anissimov, Michael. 2011. Anthropomorphism and moral realism in advanced artificial intelligence. Paper presented at the 2011 Society for Philosophy and Technology, Denton, TX, May 26–29.

Barkow, Jerome H., Leda Cosmides, and John Tooby, eds. 1992. *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Batson, Charles Daniel. 2010. *Altruism in humans.* New York: Oxford University Press.

Bostrom, Nick. 2003. Ethical issues in advanced artificial intelligence. In *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence,* ed. Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute of Advanced Studies in Systems Research / Cybernetics.

———. 2006. What is a singleton? *Linguistic and Philosophical Investigations* 5 (2): 48–54.

Bostrom, Nick, and Anders Sandberg. 2009. The wisdom of nature: An evolutionary heuristic for human enhancement. In *Human enhancement,* ed. Julian Savulescu and Nick Bostrom, 375–416. New York: Oxford University Press.

Brooks, Rodney A. 1999. *Cambrian intelligence: The early history of the new AI.* Bradford Books. Cambridge, MA: MIT Press.

Brown, Donald E. 1991. *Human universals.* New York: McGraw-Hill.

———. 2004. Human universals, human nature & human culture. *Daedalus* 133 (4): 47–54. doi:10.1162/0011526042365645.

Chalmers, David John. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

Cosmides, Leda, and John Tooby. 1992. Cognitive adaptations for social exchange. In Barkow, Cosmides, and Tooby 1992, 163–228.

Dewey, Daniel. 2011. Learning what to value. In Schmidhuber, Thórisson, and Looks 2011, 309–314.

Dijkstra, Edsger W. 1984. The threats to computing science. Paper presented at the ACM 1984 South Central Regional Conference, Austin, TX, Nov. 16–18.

Duffy, Brian R. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42 (3–4): 177–190. doi:10.1016/S0921-8890(02)00374-3.

Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Forthcoming. *The singularity hypothesis: A scientific and philosophical assessment.* Berlin: Springer.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114 (4): 864–886. doi:10.1037/0033-295X.114.4.864.

Fox, Joshua, and Carl Shulman. 2010. Superintelligence does not imply benevolence. In Mainzer 2010.

Freitas, Robert A., Jr. 1979. *Xenology: An introduction to the scientific study of extraterrestrial life, intelligence, and civilization.* 1st ed. Sacramento, CA: Xenology Research Institute. http://www.xenology.info/Xeno.htm.

Gigerenzer, Gerd, and Reinhard Selten, eds. 2001. *Bounded rationality: The adaptive toolbox.* Dahlem Workshop Reports. Cambridge, MA: MIT Press.

Goertzel, Ben. 2006. *The hidden pattern: A patternist philosophy of mind.* Boco Raton, FL: BrownWalker Press.

———. 2009. The embodied communication prior: A characterization of general intelligence in the context of embodied social interaction. In *8th IEEE international conference on cognitive informatics (ICCI '09),* ed. George Baciu. Los Alamitos, CA: IEEE Computer Society. doi:10.1109/COGINF.2009.5250687.

———. 2010. Toward a formal characterization of real-world general intelligence. In *Artificial general intelligence: Proceedings of the third conference on artificial general intelligence, AGI 2010, Lugano, Switzerland, March 5–8, 2010,* ed. Eric Baum, Marcus Hutter, and Emanuel Kitzelmann, 19–24. Advances in Intelligent Systems Research 10. Amsterdam: Atlantis Press. doi:10.2991/agi.2010.17.

Goertzel, Ben, Matthew Iklé, Izabela Freire Goertzel, and Ari Heljakka. 2009. *Probabilistic logic networks: A comprehensive framework for uncertain inference.* New York: Springer. doi:10.1007/978-0-387-76872-4.

Good, Irving John. 1965. Speculations concerning the first ultraintelligent machine. In *Advances in computers,* ed. Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Graimann, Bernhard, Brendan Allison, and Gert Pfurtscheller, eds. 2010. *Brain-computer interfaces: Revolutionizing human-computer interaction.* The Frontiers Collection. Berlin: Springer. doi:10.1007/978-3-642-02091-9.

Greene, Brian. 2011. *The hidden reality: Parallel universes and the deep laws of the cosmos.* New York: Alfred A. Knopf.

Greene, Joshua D. 2002. The terrible, horrible, no good, very bad truth about morality and what to do about it. PhD diss., Princeton University. http://scholar.harvard.edu/joshuagreene/files/dissertation_0.pdf.

Griffin, Donald R. 1992. *Animal minds.* Chicago: Chicago University Press.

Hall, John Storrs. 2007. *Beyond AI: Creating the conscience of the machine.* Amherst, NY: Prometheus Books.

Hanson, Robin. 1994. If uploads come first: The crack of a future dawn. *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

Hutter, Marcus. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

Jaynes, E. T. 2003. *Probability theory: The logic of science.* Ed. G. Larry Bretthorst. New York: Cambridge University Press. doi:10.2277/0521592712.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Karlsson, Fred. 2010. Syntactic recursion and iteration. In *Recursion and human language,* ed. Harry van der Hulst. Studies in Generative Grammar 104. New York: De Gruyter Mouton. doi:10.1515/9783110219258.43.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: Chicago University Press.

Landauer, R. 1961. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* 5 (3): 183–191. doi:10.1147/rd.53.0183.

Legg, Shane. 2006. Is there an elegant universal theory of prediction? In *Algorithmic learning theory: 17th international conference, ALT 2006, Barcelona, Spain, October 7–10, 2006. Proceedings,* ed. José L. Balcázar, Philip M. Long, and Frank Stephan. Lecture Notes in Computer Science 4264. Berlin: Springer. doi:10.1007/11894841_23.

———. 2008. Machine super intelligence. PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

Li, Ming, and Paul M. B. Vitányi. 1993. *An introduction to Kolmogorov complexity and its applications.* 1st ed. New York: Springer.

Lloyd, Seth. 2000. Ultimate physical limits to computation. *Nature* 406 (6799): 1047–1054. doi:10.1038/35023282.

———. 2002. Computational capacity of the universe. *Physical Review Letters* 88 (23): 237901. doi:10.1103/PhysRevLett.88.237901.

Mainzer, Klaus, ed. 2010. *ECAP10: VIII European Conference on Computing and Philosophy.* Munich: Verlag Dr. Hut.

Moravec, Hans P. 1998. When will computer hardware match the human brain? *Journal of Evolution and Technology* 1. http://www.transhumanist.com/volume1/moravec.htm.

Muehlhauser, Luke, and Louie Helm. Forthcoming. The singularity and machine ethics. In Eden, Søraker, Moor, and Steinhart, forthcoming.

Muehlhauser, Luke, and Anna Salamon. Forthcoming. Intelligence explosion: Evidence and import. In Eden, Søraker, Moor, and Steinhart, forthcoming.

Neisser, Ulric, Gwyneth Boodoo, Thomas J. Bouchard Jr., A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, et al. 1996. Intelligence: Knowns and unknowns. *American Psychologist* 51 (2): 77–101. doi:10.1037/0003-066X.51.2.77.

Newell, Allen, and Herbert Alexander Simon. 1972. *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Omohundro, Stephen M. 2008. The basic AI drives. In *Artificial general intelligence 2008: Proceedings of the first AGI conference,* ed. Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS Press.

Orseau, Laurent, and Mark Ring. 2011. Self-modification and mortality in artificial agents. In Schmidhuber, Thórisson, and Looks 2011, 1–10.

Reich, Peter A. 1972. The finiteness of natural language. In *Structuralist; selected readings,* ed. Fred W. Householder. Syntactic Theory 1. Harmondsworth, UK: Penguin.

Roach, Eleanor. 1978. Principles of categorization. In *Cognition and categorization,* ed. Eleanor Roach and Barbara B. Lloyd. Hillsdale, NJ: Lawrence Erlbaum Associates.

Salamon, Anna. 2009. Shaping the intelligence explosion. Paper presented at Singularity Summit 2009, New York, Oct. 3–4. http://vimeo.com/7318055.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole brain emulation: A roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Schmidhuber, Jürgen, Kristinn R. Thórisson, and Moshe Looks, eds. 2011. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings.* Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2.

Sotala, Kaj. 2010. From mostly harmless to civilization-threatening: Pathways to dangerous artificial intelligences. In Mainzer 2010.

———. Forthcoming. Advantages of artificial intelligences, uploads, and digital minds. *International Journal of Machine Consciousness* 4. Preprint at, http://www.xuenay.net/Papers/DigitalAdvantages.pdf.

Strannegård, Claes. 2005. Anthropomorphic artificial intelligence. In *Kapten mnemos kolumbarium,* 169–181. Filosofiska meddanden - Webbserien 33. Department of Philosophy, University of Gothenburg, Sweden. http://www.phil.gu.se/posters/festskrift2/mnemo_strannegard.pdf.

Tarleton, Nick. 2010. *Coherent extrapolated volition: A meta-level approach to machine ethics.* The Singularity Institute, San Francisco, CA. http://singinst.org/upload/coherent-extrapolated-volition.pdf.

Tegmark, Max. 2004. Parallel universes. In *Science and ultimate reality: Quantum theory, cosmology, and complexity,* ed. John D. Barrow, Paul C. W. Davies, and Charles L. Harper Jr., 459–491. New York: Cambridge University Press.

Tenenbaum, Joshua B., Thomas L. Griffiths, and Charles Kemp. 2006. Theory-based Bayesian models of inductive learning and reasoning. In Probabilistic models of cognition. Special issue, *Trends in Cognitive Sciences* 10 (7): 309–318. doi:10.1016/j.tics.2006.05.009.

Tooby, John, and Leda Cosmides. 1992. The psychological foundations of culture. In Barkow, Cosmides, and Tooby 1992, 19–136.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59 (236): 433–460. doi:10.1093/mind/LIX.236.433.

Veness, Joel, Kee Siong Ng, Marcus Hutter, William Uther, and David Silver. 2011. A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research* 40:95–142. doi:10.1613/jair.3125.

Vidal, Jacques J. 1973. Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering* 2:157–180. doi:10.1146/annurev.bb.02.060173.001105.

Voss, Peter. 2007. Essentials of general intelligence: The direct path to artificial general intelligence. In *Artificial general intelligence,* ed. Ben Goertzel and Cassio Pennachin, 131–157. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_4.

Yampolskiy, Roman V. 2011a. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI 2011), Thessaloniki, Greece, Oct. 3–4.

———. 2011b. What to do with the singularity paradox? Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI 2011), Thessaloniki, Greece, Oct. 3–4.

———. 2012a. AI-complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system. *ISRN Artificial Intelligence* 2012:271878. doi:10.5402/2012/271878.

———. 2012b. Leakproofing the singularity: artificial intelligence confinement problem. *Journal of Consciousness Studies* 2012 (1–2): 194–214. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00014.

———. Forthcoming. Turing test as a defining feature of AI-completeness. In *Artificial intelligence, evolutionary computing and metaheuristics: In the footsteps of Alan Turing*, ed. Xin-She Yang. Studies in Computational Intelligence 427. Berlin: Springer.

Yampolskiy, Roman V., and Joshua Fox. Forthcoming. Safety engineering for artificial general intelligence. *Topoi.* Preprint at, http://joshuafox.com/media/YampolskiyFox__SafetyEngineeringforAGI.pdf.

Yudkowsky, Eliezer. 2003. Foundations of order. Paper presented at the 2003 Foresight Senior Associates Gathering. http://singinst.org/upload/foresight.pdf.

———. 2004. *Coherent extrapolated volition.* The Singularity Institute, San Francisco, CA, May. http://singinst.org/upload/CEV.html.

———. 2006. The human importance of the intelligence explosion. Paper presented at Singularity Summit 2006, Stanford, CA, May 13. http://singinst.org/upload/singularitysummit.pdf.

———. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks,* ed. Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

———. 2011. Complex value systems in Friendly AI. In Schmidhuber, Thórisson, and Looks 2011, 388–393.