International Journal of Machine Consciousness


# ADVANTAGES OF ARTIFICIAL INTELLIGENCES, UPLOADS, AND DIGITAL MINDS

KAJ SOTALA
*Department of Computer Science, University of Helsinki*
*Helsinki, Finland*
*kaj.sotala@helsinki.fi*

Received Day Month Year
Revised Day Month Year

I survey four categories of factors that might give a digital mind, such as an upload or an artificial general intelligence, an advantage over humans. Hardware advantages include greater serial speeds and greater parallel speeds. Self-improvement advantages include improvement of algorithms, design of new mental modules, and modification of motivational system. Co-operative advantages include copyability, perfect co-operation, improved communication, and transfer of skills. Human handicaps include computational limitations and faulty heuristics, human-centric biases, and socially motivated cognition. The shape of hardware growth curves, as well as the ease of modifying minds, are found to have a major impact on how quickly a digital mind may take advantage of these factors.

*Keywords*: artificial general intelligence; hard takeoff; mind uploading; technological singularity; whole brain emulation

## 1. Introduction

A digital mind is a mind that runs on a computer. One type of a digital mind is a mind upload, a hypothetical mind that was originally human, but that has been moved into a digital format and is being run as a software program on a computer. Another type of a digital mind is that of an artificial general intelligence (AGI). While uploads are based on taking existing human minds and closely replicating them in software, AGI may be built on computer science principles and have little or no resemblance to the human psyche.

Either type of a digital mind might be created within a timeframe of decades to centuries. A recent roadmap charting the technological requirements for creating uploads suggests that they may be feasible by mid-century [Sandberg & Bostrom, 2008]. Sotala & Valpola [this issue] note that research into prostheses replicating the functions of the hippocampus and the cerebellum is well under way, and suggest that a feasible future development would be "exocortices", implants that can be connected to human brains and which gradually take over cortical brain function.

Interest in AGI research is also growing, with an increasing number of special sessions, workshops, and conferences devoted specifically to topics such as AGI having been held in the recent years [Baum *et al.*, 2011]. In an expert assessment survey conducted at the Artificial General Intelligence 2009 (AGI-09) conference, the median estimates for when there would be a 10%, 50%, or a 90% chance of having an AGI capable of passing the third grade were 2020, 2030 and 2075, respectively [Baum *et al.*, 2011]. In an informal survey conducted at the 2011 Winter Intelligence Conference, the median estimates for when there would be a 10%, 50%, or a 90% chance of developing human-level machine intelligence were 2028, 2050, and 2150 [Sandberg & Bostrom, 2011].

There has been some previous work focused on examining the consequences of creating digital intelligences. Focusing specifically on uploads, researchers have examined some of the economical consequences of the ability to copy minds [Hanson, 1994, 2008], as well as the improved coordination ability stemming from being able to copy, delete and restore minds [Shulman, 2010]. Hanson [2001] and Kaas *et al.* [2010] look more generally at the economic effects of digital minds that can be copied. Sotala & Valpola [this issue] explore the way that mind uploading may lead to "mind coalescence", the ability to merge previously separate minds together.

Other researchers have argued that once we have AGI, it will surpass the capabilities of humans in many domains at drastic speed (see e.g. [Vinge, 1993; Bostrom, 2000; Yudkowsky, 2008a; Chalmers, 2010]). A "hard takeoff" [Yudkowsky, 2001a; Bugaj & Goertzel, 2007; Hall, 2008; Vinge, 2008] involves an AGI reaching a point which allows it to quickly accumulate various advantages and influence, becoming a dominant power before humans have the time to properly react. Should existing human preparation be inadequate to such a drastic event, there could be serious consequences, up to and including human extinction [Bostrom, 2002; Yudkowsky, 2008a; Chalmers, 2010]. This paper attempts to study the consequences of creating a digital mind in terms of the advantages that they may enjoy over humans.

### 1.1. *Intelligence, optimization power, and advantages*

A great variety of definitions have been offered for the word "intelligence". A definition which seems to summarize the essential content in most of them is that intelligence measures an agent's ability to achieve goals in a wide range of environments [Legg & Hutter, 2007]. A mind has goals which it tries to achieve, and more intelligent minds are better at finding, inventing, and evaluating various ways of achieving their goals. A generalization of the concept of intelligence is the notion of optimization power [Yudkowsky, 2008a; Muehlhauser & Helm, 2011], an agent's general ability to achieve its goals. While intelligence is derived from what are generally considered "mental" faculties, an agent's optimization power is also a factor of things such as its allies and resources, as well as its ability to obtain more of them.

The crucial risk involved in the creation of digital minds is the possibility of

creating minds whose goals are very different from humanity's, and who end up possessing more optimization power than humanity does. Current human preferences and desires seem to be very complex and not well-understood: there is a strong possibility that only a very narrow subset of all possible goals will, if successful, lead to consequences that would be considered favorable by humans [Yudkowsky, 2008a; Muehlhauser & Helm, 2011]. If digital intelligences are created and end up having (as a group) more optimization power than humanity does, and their goals are very different from humanity's goals, then the consequences are likely to be considered very bad by most humans.

This paper attempts to analyze the consequences of creating a digital mind from the perspective of the optimization power that they might accumulate. Factors that may lead to digital minds accumulating more optimization power than humans are called advantages; factors that may lead to digital minds accumulating less optimization power than humans are called disadvantages.

## 2. Hardware advantages

A digital mind running on a computer system can upgrade the system to utilize more powerful hardware, while biological humans cannot drastically upgrade their brains. Suppose that there is some minimum hardware configuration that provides a digital mind with roughly the same processing power and memory as a human brain. Any increase in hardware resources past this point is a hardware advantage in favor of the digital mind.

### 2.1. *Superior processing power*

The amount of processing power required to run a digital mind is currently unknown[a]. For uploads, Sandberg & Bostrom [2008] place $10^{18}$ to $10^{25}$ FLOPS as the most likely amount required to run one in real time. They estimate that current trends would make these levels available for purchase at the cost of 1 million dollars around 2019 to 2044.

An upload attempts to accurately emulate the entirety of human brain function, with all relevant details intact. In contrast, AGI designers are free to use any working algorithm, regardless of its biological plausibility. The human brain is evolved to function in a way suited to the constraints of biology, which may be very different from what would run efficiently on a computer. This allows for the possibility that an AGI would require much less processing power than an upload. Estimates of the amount of processing power required to run a mind range from modern-day

[a]Some sources provide their estimates in terms of MIPS (Millions of Instructions Per Second), while others use FLOPS (Floating-Point Operations per Second). These are not directly comparable, and there is no reliable way to convert between the two. For this paper, we have used the rough estimate in Sandberg & Bostrom [2008] that FLOPS grow as MIPS to the power of 0.8. The authors warn that this trend may change, with the exponent possibly becoming larger than 1

computers [Hall, 2007] to $10^{11}$ FLOPS [Moravec, 1998] and $10^{14}$ FLOPS [Bostrom, 1997].

#### 2.1.1. *Superior serial power*

Humans perceive the world on a particular characteristic time scale. A mind being executed on a system with greatly superior serial power could run on a faster timescale than we do. For instance, a mind with twice the serial power of the human brain might experience the equivalent of two seconds passing for each second that we did, thinking twice the amount of thoughts in the same time. This advantage would be especially noticable in time-critical decision-making.

Even a small advantage would accumulate given enough time. Over the course of a year, a 10% difference in speed would give the faster mind more than an extra month. This would allow it to outcompete any mind with equal skills and resources but without the speed advantage.

In the "speed explosion" scenario [Solomonoff, 1985; Yudkowsky, 2001b; Chalmers, 2010], digital researchers, running at an accelerated speed, work to develop faster computers. If the minds doing the research could take advantage of the faster hardware they produced, the time required to develop the next generation of hardware could keep getting shorter as the researchers would be getting more done in the same time. This could continue until some bottleneck, such as the time needed to physically build the computers, or fundamental barrier was reached.

#### 2.1.2. *Increased parallel power, increased memory*

Recent advances in computing power have been increasingly parallel instead of serial. If the trend keeps up, a future computer may not be able to use its superior processing power to gain a direct increase in speed over the human brain, if the tasks in question do not parallelize well. Amdahl's law [Amdahl, 1967] states that the if a fraction f of a program's performance can be parallelized, then the speedup given by n processors instead of one is 1 / (1 - f) + f / n. A difference of several orders of magnitude in computing power might translate to a much more modest change in speed. Gustafson [1988] notes that in practice, it is the parallelizable part of a problem that grows as data is added, and the serial part remains constant. Even if increasing the number of processors didn't allow a problem to be solved in less time, it can allow a larger problem to be solved in better detail.

As the human brain works in a massively parallel fashion, at least some highly parallel algorithms must be involved with general intelligence. Extra parallel power might then not allow for a direct improvement in speed, but it could provide something like a greater working memory equivalent. More trains of thought could be pursued at once, and more things could be taken into account when considering a decision. Brain size seems to correlate with intelligence within rats [Anderson, 1993], humans [McDaniel, 2005], and across species [Deaner *et al.*, 2007], suggesting that

increased parallel power could make a mind generally more intelligent.

## 3. Self-improvement advantages

A digital mind with access to its source code may directly modify the way it thinks, or create a modified version of itself. In order to do so, the mind must understand its own architecture well enough to know what modifications are sensible. An AGI can intentionally be built in a manner that is easy to understand and modify, and may even read its own design documents. Things may be harder for uploads, especially if the human brain is not yet fully understood by the time uploading becomes possible.

Either type of mind could experiment with a large number of possible interventions, creating thousands or even millions of copies of itself to see what kinds of effects various modifications have. While some of the modifications could produce unseen long-term problems, each copy could be subjected to various intensive tests over an extended period of time to estimate the effects of the modifications. Copies with harmful or neutral modifications could be deleted, making room for alternative ones. [Shulman, 2010] Less experimental approaches might involve formal proofs of the effects of the changes to be made.

Recursive self-improvement [Yudkowsky, 2008a; Chalmers, 2010] is a situation in which a mind modifies itself, which then makes it capable of further improving itself. For instance, an AGI might improve its pattern-recognition capabilities, which would then allow it to notice inefficiencies in itself. Correcting these inefficiencies would free up processing time and allow the AGI to notice more things that could be improved.

To a limited extent, humans have been engaging in recursive self-improvement as the development of new technologies and forms of social organization has made it possible to organize better and develop yet more advanced technologies. Yet the core of the human brain has remained the same. If changes could be found that sparked off more changes, which kept sparking off more changes, the result could be a greatly improved form of intelligence. [Yudkowsky, 2008a]

### 3.1. *Improving algorithms*

A digital mind could come across algorithms in itself that could be be improved. For instance, they could be made faster, to consume less memory, or to rely on fewer assumptions. In the simplest case, an AGI implementing some standard algorithm might come across a paper detailing an improved implementation of it. Then the old implementation could be simply replaced with the new one. An upload with emulated neurons might alter itself so as to mimic the effects of drugs, neurosurgery, genetic engineering and other interventions [Shulman, 2010].

In the past, improvements in algorithms have sometimes been even more important than improvements in hardware. The President's Council [2010] mentions that performance on a benchmark production planning model improved by a factor of 43 million between 1988 and 2003. Out of the improvement, a factor of roughly 1,000

was due to better hardware and a factor of roughly 43,000 was due to improvements in algorithms. Also mentioned is an algorithmic improvement of roughly 30,000 for mixed integer programming between 1991 and 2008.

### 3.2. *Designing new mental modules*

A mental module, in the sense of functional specialization [Cosmides & Tooby, 1994; Barrett & Kurzban, 2006], is a part of a mind that specializes in processing a certain kind of information. Specialized modules are much more effective than general-purpose ones, for the number of possible solutions to a problem in the general case is infinite. Research in a variety of fields, including artificial intelligence, developmental psychology, linguistics, perception and semantics has shown that a system must be predisposed to processing information within the domain in the right way or it will be lost in the sea of possibilities. [Tooby & Cosmides, 1992] Many problems within computer science are intractable in the general case, but can be efficiently solved by algorithms customized for specific special cases with useful properties that are not present in general [Cormen *et al.*, 2009]. Correspondingly, many specialized modules have been proposed for humans, including modules for cheater-detection, disgust, face recognition, fear, intuitive mechanics, jealousy, kin detection, language, number, spatial orientation and theory of mind [Barrett & Kurzban, 2006].

Specialization leads to efficiency: to the extent that regularities appear in a problem, an efficient solution to the problem will exploit those regularities [Kurzban, 2010]. A mind capable of modifying itself and designing new modules customized for specific tasks might eventually outperform biological minds in any domain, even presuming no hardware advantages. In particular, any improvements in a module specialized for creating new modules would have a disproportionate effect.

It is important to understand what specialization means in this context, for it is frequently misunderstood. For instance, Bolhuis *et al.* [2011] argue against functional specialization in nature by citing examples of "domain-general learning rules" in animals. However, Barrett & Kurzban [2006] point out that even seemingly domain-general rules, such as the modus ponens rule of formal logic, operate in a restricted domain: representations in the form of if-then statements. In defining the domain of a module, what matters is not the content of the domain, but the formal properties of the processed information and the computational operations performed on the information. Positing functional modules in humans also does not imply genetic determination, nor that the modules could necessarily be localized to a specific part of the brain. [Barrett & Kurzban, 2006]

A special case of a new mental module is the design of a new sensory modality, such as that of vision or hearing. Yudkowsky [2007] discusses the notion of new modalities, and considers the detection and identification of invariants to be one of the defining features of a modality. In vision, changes in lightning conditions may entirely change the wavelength of light that is reflected off a blue object, but it

is still perceived as blue. The sensory modality of vision is then concerned with, among other things, extracting the invariant features that allow an object to be recognized as being of a specific color even under varying lightning.

Brooks [1987] mentions invisibility as an essential difficulty in software engineering. Software cannot be visualized in the same way physical products can be, and any visualization can only cover a small part of the software product. Yudkowsky [2007] suggests a codic cortex designed to model code the same way that the human visual cortex is evolved to model the world around us. Whereas the designer of a visual cortex might ask "what features need to be extracted to perceive both an object illuminated by yellow light and an object illuminated by red light as 'the color blue'?", the designer of a codic cortex might ask "what features need to be extracted to perceive the recursive algorithm for the Fibonacci sequence and the iterative algorithm for the Fibonacci sequence as 'the same piece of code'?" Speculatively, new sensory modalities could be designed for various domains for which existing human modalities are not optimally suited.

### 3.3. *Modifiable motivation systems*

Humans frequently suffer from problems such as procrastination, boredom, mental fatigue, and burnout. A mind which did not become bored or tired with its work would have a clear advantage over humans. Shulman [2010] notes two ways by which uploads could overcome these problems. Uploads could be copied while they were in a rested and motivated state. When they began to tire, they could be deleted and replaced with the "snapshot" taken while they were still rested. Alternatively, their brain state could be edited so as to eliminate the neural effects of boredom and fatigue. An AGI might not need to have boredom built into it in the first place.

The ability for a mind to modify its own motivational systems also has its own risks. Wireheading [Yudkowsky, 2001a; Omohundro, 2008] is a phenomenon where a mind self-modifies to make it seem like it is achieving its goals, even though it is not. For instance, an upload might try to eliminate its stress about its friends dying by creating a delusion about them always being alive. Once a mind has wireheaded, it may no longer want to fix its broken state.

Even if wireheading-related problems were avoided, a mind altering its own motivations still risks an outcome where its ability to pursue its original goals is worsened. To avoid such problems, a mind might attempt to formally prove that proposed changes do not alter its current goals [Yudkowsky, 2008a], or it may produce modified copies of itself and subject the copies to an intesive testing regimen [Shulman, 2010].

## 4. Co-operative advantages

### 4.1. *Copyability*

A human child takes nine months to gestate, after which close to two or even three decades are typically needed before it can do productive work, depending on the

type of work. Raising a child is expensive, costing on average between 216,000 and 252,000 dollars over 18 years in the United States [Lino, 2010]. In contrast, a digital mind can be copied very quickly and doing so has no cost other than access to the hardware required to run it. Hanson [1994, 2008] estimates that copyable workers could rapidly come to dominate major portions of the economy, as the mind willing to work for the lowest wage could be copied until there was no more demand for workers of that type. Although such workers would be individually poor, they would control a large amount of wealth as a group. Whether they could take advantage of it would depend on their ability to co-operate.

For human populations, the maximum size of the population depends primarily on the availability of resources such as food and medicine. For populations of digital minds, the maximum size of the population depends primarily on the amount of hardware available. A digital mind could obtain more hardware resources by legitimitely buying them, or by employing illegal means such as hacking and malware.

Modern-day botnets are networks of computers that have been compromised by outside attackers and are used for illegitimate purposes. Estimates of their size range from one study saying the effective sizes of botnets rarely exceed a few thousand bots, to a study saying that botnet sizes can reach 350,000 members [Rajab *et al.*, 2007]. Currently, the distributed computing project Folding@Home, with 290,000 active clients, can reach speeds in the $10^{15}$ FLOPS range [Pande Lab, 2012]. A relatively conservative estimate that presumed a digital mind couldn't hack into more computers than the best malware practicioners of today, that a personal computer would have a hundred times the computing power of today, and that the mind took $10^{13}$ FLOPS to run, would suggest that a single mind could spawn 12,000 copies of itself.

Resorting to illegal means may not be necessary if digital minds are allowed to own property, or at least earn money. Uploads might be legally considered the same person as before the upload. AGIs owned by a company or private individual can accumulate property for their owner, or they may try to set up a front company to act through. Creating new copies of a mind is profitable until the costs of maintaining a new copy exceed the profits that copy is capable of generating, so copies might be created until the wage a copy can earn falls to the level of maintaining the hardware [Hanson, 1994, 2008]. Hanson [2008] argues that copying would drive wages down to machine-subsistence levels, leading to "insectlike urban densities, with many billions [of digital minds] living in the volume of a current skyscraper, paying astronomical rents that would exclude most humans".

For uploads, who might not be capable of co-operating with each other any better than current-day humans do, this might be a disadvantage rather than an advantage. But if they could, or if an AGI spawned many copies of itself, then the group could pool their resources and together control a large fraction of the wealth in the world.

### 4.2. *Perfect co-operation*

Human capability for co-operation is limited by the fact that humans have their own interests in addition to those of the group. Olson [1965] showed that it is difficult for large groups of rational, selfish agents to effectively co-operate even to achieve common goals, if those goals can be characterized as obtaining a public good for the group. Public goods are those that, once obtained, benefit everyone and cannot effectively be denied to anyone. Because an individual's contribution has a neglible effect on the achievement of the good, all group members have an incentive to free ride on the effort.

One implication is that smaller groups often have an advantage over larger ones. For one, co-operation is easier to enforce in a smaller group. Additionally, it is may be beneficial for e.g. a large company to lobby for laws benefiting the whole industry, since it is large enough to benefit from the laws even if it had to shoulder almost all of the costs of lobbying itself. Smaller companies forgo such lobbying and free ride on the large company's investment. This leads to lobbying investment that is suboptimal for the industry as a whole. [Olson, 1965; Mueller, 2003] Self-interest is a natural consequence of evolution, as it increases the odds that an organism survives long enough to breed. Drives such as self-preservation are also natural instrumental values for any intelligent agent, for only entities that continue to exist can work towards their goals [Omohundro, 2008].

However, minds might be constructed to lack any self-interest, particularly if multiple copies of the same mind existed and the destruction of one would not seriously threaten the overall purpose. Such entities minds could be identical to one another, share the same goal system, and co-operate perfectly with one another with no costs from defection or from systems for enforcing co-operation.

In the case of uploads, this could happen through copying an upload in suitable ways, creating a "superorganism". An upload that has been copied may hold the view that being deleted is an acceptable cost to pay, for as long as other copies of it survive. Uploads holding this view would then be ready to make large sacrifices for the rest of the superorganism, and might employ various psychological techniques to reinforce this bond. [Shulman, 2010] Minds wishing to work for a common goal might also choose to connect their brains together, more or less coalescing together into a single mind [Sotala & Valpola, this issue]. A lighter form of mind coalescence might also be used to strenghten the unity of a superorganism.

### 4.3. *Superior communication*

Misunderstandings are notoriously common among humans. AGI could potentially need to spend much less effort on communication. Language can be thought of as symbols that map to different individual's conceptual spaces, with miscommunication occurring because of different mappings [Honkela *et al.*, 2008]. Efficiency of communication could be improved by having very similar conceptual spaces (aiding communication between copies) or via custom-tailoring mental modules to the do-

main of conceptual mapping. Such modules could e.g. simulate the interpretations to different messages emerging from a wide variety of conceptual spaces and seek to include the caveats excluding those interpretations, or directly communicate parts of the presumed conceptual spaces using some standard language.

Humans are limited as to the speed at which they can listen to others, or read written text, without loss of comprehension. Increased skill at doing conceptual mapping, as well as increased processing power, could plausibly increase these rates. Digital minds could also communicate at higher bandwidths, transmitting a vastly larger amount of information at once. If necessary, minds could join their minds together, exchanging thoughts directly [Sotala & Valpola, this issue].

### 4.4. *Transfer of skills*

Copying parts of a mind is a special case of copying a whole mind. To the extent that skills can be modularized, digital minds could create self-contained skill modules to be shared with others. In the most extreme case, a population of minds could outsource all of their skills to a very small number of cognitive modules, only learning a small number of things themselves [Bostrom, 2004]. Whenever one mind had mastered a skill, it could share it with all the others.

## 5. Human handicaps

Humans frequently reason in suboptimal or incorrect ways (see e.g. Stanovich [2009] or Kahneman [2011]). Such failures of reasoning have an enormous negative impact on society. Among other things, they cause people to suffer from a worse standard of living, make bad investments, become more easily manipulated, end up falsely accused or imprisoned by the authorities, increase the mortality rate, or even fall prey to scams serious enough to crash a national economy [Stanovich, 2009]. A mind that was immune to such biases would reason more reliably than we do, while possibly exploiting our biases. An upload may attempt to self-modify to overcome its biases, while an AGI might never have the biases in the first place.

### 5.1. *Biases from computational limitations or false assumptions*

Some human biases can be seen as assumptions or heuristics that fail to reason correctly in a modern environment, or as satisficing algorithms that do the best possible job given human computational resources [Gigerenzer & Brighton, 2009]. A digital mind could potentially overcome most if not all of the biases that plague human reasoning, either by rewriting its algorithms to better suit the environment or to better take into account growing computational resources.

For example, when faced with a difficult question, human brains have a tendency to instead solve an easier question and treat it as the answer to the more difficult question. If asked "how much will this company have grown in five years", an intuitive answer might be generated based on the question "how fast has the

company been growing so far". These kinds of heuristics often function well, but they sometimes fail to take into account crucial factors, such as reasons for why the company might fail to keep up its historical growth rates. While such factors can be taken into account by explicitly thinking about them, it requires explicit thought and is not done automatically. [Kahneman, 2011] A digital mind might be capable of editing the heuristics it uses for answering such questions, and avoid the risk of accepting the heuristic answer as a fact without further evaluation, like humans often do.

One's susceptibility to some major biases, such as overconfidence and hindsight bias, correlates negatively with one's general intelligence. This suggests that computational limitations cause at least some of the flaws in human reasoning [Stanovich & West, 1998].

### 5.2. *Human-centric biases*

People tend to think of the capabilities of non-human minds, such as God or an artificial intelligence, as if the minds in question were human. This tendency persists even if humans are explicitly instructed to act otherwise. [Barrett & Keil, 1996] This is a special case of biases due to false assumptions.

Evolutionarily, other minds have constituted possibly the single most important selection pressure facing any single human - the extent to which one can co-operate with others and avoid being exploited will to a large extent determine one's success in life. Because mental states such as beliefs, motives, intentions and emotions cannot be directly observed and have to be inferred, we are likely to have evolved a large amount of algorithms and modules for inferring such states on the basis of very subtle cues. [Cosmides & Tooby, 1994]. Since we have never had to model the thoughts of non-human minds to this extent, these modules will automatically try to model any minds we're dealing with using the same principles. Thus they will carry over human-centric assumptions when we attempt to model the behavior of non-human minds [Yudkowsky, 2008a].

To some extent, the modules may even assume we're dealing with humans similar to ourselves: neural systems we use for modeling others overlap with those used for processing information about ourselves [Uddin *et al.*, 2007]. Humans in a "cold" (non-emotional, non-aroused) state frequently overestimate their degree of self-control in a "hot" (emotional, aroused) state [Loewenstein, 2005]. Even a relatively mild difference between oneself and the mind to be modeled can thus lead to erronous predictions.

To the extent that digital minds reason and behave unlike humans, our attempts to intuitively predict their behavior will be based on false premises. To some extent, this disadvantage may be symmetrical, in that e.g. uploads engaged in self-modification will fail to understand how non-modified humans would behave. AGIs may not start out with any efficient models for human behavior in the first place. Over time, biological humans will accumulate expertise in predicting digital

minds and digital minds may learn and self-modify to better understand humans. However, biological humans may be at a disadvantage if self-modification is easy, for the assumptions applying to digital minds may change faster than humans can keep up with.

### 5.3. *Biases from socially motivated cognition*

It has also been proposed that humans have evolved to acquire beliefs which are socially beneficial, even if those beliefs weren't true. [Trivers, 2000; Kurzban & Aktipis, 2007; Kurzban, 2010]. For instance, minds may be biased to believe that they'll be successful in order to persuade others to ally with them. Even more seriously, minds may not be built to actually question the accuracy of their beliefs, but to rationalize reasonable-sounding explanations for their initial emotional reaction to a concept. If human reasoning is strongly enough biased to come up with popular or self-beneficial theories, instead of theories that are actually true, a mind without such a bias could be immensly more effective at reaching the correct theories.

## 6. Discussion

Two main questions seem to emerge:

**What do hardware growth curves look like?** A number of advantages are either completely based on improved hardware (superior processing power) or made greatly stronger by it (overcoming biases, designing new mental modules, copyability, superior communication). Therefore the faster hardware advances, the steeper the takeoff.

Digital minds are subject to the risk of hardware overhang [Yudkowsky, 2008b; Shulman & Sandberg, 2010]. If software development proceeds slower than hardware development, then the hardware required for digital minds may be available far earlier than the software. When the software for digital minds is developed, the minds could have at their disposal much more hardware resources than is strictly necessary to run them, giving them an unexpected advantage.

Even if hardware development stopped for a while and digital minds were stuck on a certain level that put them on roughly equal footing with humanity, this situation could not be relied upon to persist. Any future breakthrough in hardware would have the potential to upset the situation and give the digital minds a decisive advantage. Lloyd [2000] estimates the ultimate physical limits on hardware to allow for a one-liter, one-kilogram computer capable of carrying out $10^{50}$ OPS if we allow for computers made of exotic matter that explode during the calculation, or $10^{40}$ OPS if we constrain ourselves to computers made of ordinary matter. Whatever estimate we use for the human brain's processing power, we cannot with any certainty presume that it would be near the achievable physical limits on computation.

**How modifiable are minds?** A number of advantages (improved parallel power, overcoming biases, designing new mental modules, perfect co-operation, superior communication, transfer of skills) rely to differing degrees on the assumption

that digital minds are easy to understand and modify. To the degree to which this assumption is untrue, these advantages become less pronounced. Loosemore [2007] argues that there may be a disconnect between the local behavior of interacting elements and the global behavior of the system in a mind, so that generally intelligent behavior might not be derivable from mathematical rules. This would reduce the pace of self-improvement, though self-improvement would still be possible via systematic exploration of related mental architectures.

A closely related and important question is the "intelligibility of intelligence" [Salamon *et al.*, 2010] - the question of whether the core of general intelligence could be expressed in a compact intelligible theory, like the theory of relativity, or whether it is more akin to a "swiss army knife" of incremental solutions and special tricks. If intelligence is generally unintelligible and hard, then improving upon minds might prove difficult and slow.

Bach [2010] argues that like AGIs, human organizations such as corporations, administrative and governmental bodies, churches and universities are intelligent agents that are more powerful than individual humans, and that the development of AGI would increase the power of organizations in a quantitative way but not cause a qualitative change.

Humans grouping into organizations are to some degree capable of taking advantage of increased parallel (but not serial) speed by adding more individuals. While organizations can institute guidelines such as peer review that help combat bias, working in an organization can introduce biases of its own, such as groupthink [Esser, 1998]. They cannot design new mental modules or benefit from any of the co-operative advantages digital minds may enjoy. Possibly their largest shortcoming is their reduced efficiency as the size of the organization grows and their general susceptibility to having their original goals hijacked by smaller interest groups within the organization [Olson, 1965].

## 7. Conclusions

Digital minds potentially enjoy a number of advantages, all of which might make it easier for them to succeed in their goals. Hardware improvements could allow digital minds to think faster and to consider more things at once. Self-improvement advantages would allow digital minds to modify themselves, possibly in a recursive fashion where initial improvements kept sparking off further improvements. Algorithms could be improved to give an even larger boost than hardware improvements could, new mental modules could be designed for new kinds of domains, and the various motivational systems plaguing humans could be overcome. Co-operative advantages include copyability, which would let a mind replicate itself many times over, and the potential for perfect co-operation would eliminate conflict between copies of the same mind. Superior communication and the ability to transfer skills would further help matters. Finally, humans may suffer from various handicaps in their mental architecture. They might employ inadequate heuristics, have a harder

time modeling digital minds than digital minds have modeling humans, and suffer from socially motivated cognition.

The extent to which these potential advantages could be realized is an open question. Hardware advantages rely on hardware progressing, copyability relies on there being enough hardware to run a large number of minds, and nearly all of the others rely on minds being sufficiently modifiable.

It is possible that hard takeoff scenarios have gotten a disproportionate amount of attention in discussions of digital mind advantages. From a safety viewpoint, assuming that a digital mind can bootstrap itself to superintelligence in a matter of weeks or hours is a conservative guess, in the sense that it's the scenario that leaves others the least time to prepare. [Yudkowsky, 2001a, 2008a] We should try to anticipate such a scenario, because if we do not, there will be little or no time to react when it happens. Yet debates over the plausibility of a hard takeoff distract from the fact that digital minds developing over a timespan of years or decades is a dangerous scenario as well. Digital minds might be developed, then be relatively weak for an extended time, until some hardware or software breakthrough suddenly allowed them to become considerably more powerful. Likewise, coalitions could initially form, keeping each other in check, until something happened to break the balance. The possibility of a hard takeoff is definitely real and deserves attention, but it is far from the only danger.

## Acknowledgments

This paper began as a collaboration with Zack M. Davis, who then withdrew from the project. Regardless, this paper might not have existed without his early work.

The author would also like to thank Alexandros Marinos, Anna Salamon, Louie Helm, several pseudonymous commenters on `http://lesswrong.com`, two anonymous reviewers, and anyone else he may have forgotten but who did not deserve to be forgotten, for their feedback on the article.

## References

Amdahl. G. M. [1967] "Validity of the single-processor approach to achieving large scale computing capabilities", in *AFIPS Conference Proceedings, vol. 30* (Atlantic City, N.J..), pp. 483-485.

Anderson B. [1993] Evidence from the rat for a general factor that underlies cognitive performance and that relates to brain size: Intelligence? *Neuroscience Letters* **153**(1), 98-102.

Bach, J. [2010] "How the Singularity of Artificial Intelligence might be achieved, and why it does not matter," in *Proc. VIII European Conference on Computing and Philosophy (ECAP10)* (Mnchen, Germany), pp. 471-475.

Barrett, J. L. and Keil, F. C. [1996] Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts. *Cognitive Psychology* **31**, 219-247.

Barrett, H. C. and Kurzban, R. [2006] Modularity in Cognition: Framing the Debate. *Psychological Review*, **113**(3), 628-647.

Baum, S. D., Goertzel, B., Goertzel, T. G. [2011] "How long until human-level AI? Results from an expert assessment." *Technological Forecasting and Social Change*, **78**(1), 185-195.

Bolhuis, J. J., Brown G. R., Richardson R. C. and Laland K. N. [2011] Darwin in Mind: New Opportunities for Evolutionary Psychology. *PLoS Biol* **9**(7): e1001109. doi:10.1371/journal.pbio.1001109

Bostrom, N. [1997] How long before superintelligence? *Int. Jour. of Future Studies*, **2**.

Bostrom, N. [2000]. When machines outsmart humans. *Futures*, **35**(7), 759 - 764

Bostrom, N. [2002] Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards, *Journal of Evolution and Technology*, **9**.

Bostrom, N. [2004] The Future of Human Evolution, in Tandy, C. (ed.) *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing* (Ria University Press), pp. 339–371.

Bostrom, N. and Yudkowsky, E. [2011] The Ethics of Artificial Intelligence, in Ramsey, W. and Frankish, K. (eds.) *Cambridge Handbook of Artificial Intelligence* (Cambridge University Press).

Brooks, F. P. [1987] No Silver Bullet - Essence and Accidents of Software Engineering. *IEEE Computer* **20**(4), 10-19.

Bugaj, S. V. and Goertzel, B. [2007] Five Ethical Imperatives and their Implications for Human-AGI Interaction. *Dynamical Psychology.*

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

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. [2009]. *Introduction to algorithms* (The MIT press, Cambridge, Massachusetts).

Cosmides, L. and Tooby, J. [1994]. Origins of domain specificity: The evolution of functional organization, in Hirschfeld, L. and Gelman, S. (eds.), *Mapping the mind: Domain specificity in cognition and culture* (Cambridge University Press, Cambridge) pp. 85-116.

Deaner, R., Isler, K. Burkhart, J., and Schaik, C. [2007]. Overall brain size, not encephalization quotient, best predicts cognitive ability across non-human primates. *Brain, Behavior, and Evolution* **70**(2), 115-124.

Esser, J. K. [1998] Alive and Well after 25 Years: A Review of Groupthink Research. *Organizational Behavior and Human Decision Processes*, **73**(2/3), 116-141.

Gigerenzer, G. and Brighton, H. [2009] Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science* **1**, 107-143.

Gustafson, J. L. [1988] Reevaluating Amdahl's Law. *Communications of the ACM* **31**(5), 532-533.

Hall, J. S. [2007] *Beyond AI: Creating the Conscience of the Machine* (Prometheus).

Hall, J.S. [2008] "Engineering Utopia," in *Artificial general intelligence, 2008: pro-*

*ceedings of the First AGI Conference.*

Hanson, R. [1994] If uploads come first, *Extropy* **6**(2), `http://hanson.gmu.edu/uploads.html`.

Hanson, R. [2001] Economic Growth Given Machine Intelligence. `http://hanson.gmu.edu/aigrow.pdf`.

Hanson, R. [2008] Economics of the singularity, *IEEE Spectrum*, 37–42, June 2008.

Honkela, T., Knnen, V., Lindh-Knuutila, T., Paukkeri, M-S. [2008] Simulating Processes of Concept Formation and Communication. *Journal of Economic Methodology* **15**(3), 245-259.

Kaas, S., Rayhawk, S., Salamon, A., Salamon, P. [2010] "Economic Implications of Software Minds," in *Proc. VIII European Conference on Computing and Philosophy (ECAP10)* (Mnchen, Germany), pp. 431-437.

Kahneman, D. [2011] *Thinking, Fast and Slow* (Farrar, Straus and Giroux).

Kurzban, R. and Aktipis, C. A. [2007] Modularity and the Social Mind: Are Psychologists Too Self-Ish? *Personality and Social Psychology Review* **11**(2).

Kurzban, R. [2010] *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind* (Princeton University Press).

Legg, S. and Hutter, M. [2007] Universal Intelligence: A Definition of Machine Intelligence. *Minds & Machines* **17**(4), 391-444.

Lino, M. [2010] Expenditures on Children by Families, 2009. U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. Miscellaneous Publication No. 1528-2009.

Lloyd, S. [2000] Ultimate physical limits to computation. *Nature* **406**, 1047-1054.

Loewenstein, G. [2005]: Hot-cold empathy gaps and medical decision making. *Health Psychology* **24**, S49-S56.

Loosemore, R. P. W. [2007]. "Complex Systems, Artificial Intelligence and Theoretical Psychology," in Goertzel, B. and Wang, P. (eds.) *Proceedings of the 2006 AGI Workshop.*

McDaniel, M. A. [2005] Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence* **33**, 337-346.

Moravec, H. [1998] When Will Computer Hardware Match the Human Brain? *Journal of Evolution and Technology, vol. 1*, `http://www.transhumanist.com/volume1/moravec.htm`.

Muehlhauser, L. and Helm, L. [2011] Machine Ethics and the Singularity. `http://commonsenseatheism.com/wp-content/uploads/2011/11/Muehlhauser-Helm-The-Singularity-and-Machine-Ethics-draft.pdf`

Mueller, D. C. [2003] *Public Choice III* (Cambridge University Press).

Olson, M. [1965] *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, Cambridge MA).

Omohundro, S. [2008] "The basic AI drives," in Wang, P. Goertzel, P. and Franklin, S. (eds.), *Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applicaions, Volume 171* (IOS Press) 483-494.

Pande Lab [2012] Client statistics by OS. `http://fah-web.stanford.edu/cgi-bin/main.py?qtype=osstats`.

President's Council of Advisors on Science and Technology [2010] Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. *Report to the President and Congress*. `http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf`.

Rajab, M.A., Zarfoss, J., Monrose, F. and Terzis, A. [2007] "My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging," in *Proceedings of 1st Workshop on Hot Topics in Understanding Botnets (HotBots '07)*.

Sandberg, A. and Bostrom, N. [2008] Whole brain emulation: a roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University. `http://www.fhi.ox.ac.uk/reports/2008-3.pdf`.

Sandberg, A. and Bostrom, N. [2011] Machine Intelligence Survey, Technical Report #2011-1, Future of Humanity Institute, Oxford University: pp. 1-12. `http://www.fhi.ox.ac.uk/reports/2011-1.pdf`.

Salamon, A., Rayhawk, S. and Kramr, J. [2010] "How intelligible is intelligence?" in *Proc. VIII European Conference on Computing and Philosophy (ECAP10)* (Mnchen, Germany), pp. 438-443.

Shulman, C. [2010] Whole Brain Emulation and the Evolution of Superorganisms, `http://singinst.org/upload/WBE-superorganisms.pdf`.

Shulman, C. and Sandberg, A. [2010] "Implications of a software-limited singularity," in *Proc. VIII European Conference on Computing and Philosophy (ECAP10)* (Mnchen, Germany), pp. 463-470.

Solomonoff, F. [1985] The time scale of artificial intelligence: Reflections on social effects. *North-Holland Human Systems Management* **5**, 149-153.

Sotala, K. and Valpola, H. [2012] Coalescing minds: brain uploading-related group mind scenarios. *Int. J. of Machine Consciousness*.

Stanovich, K. E. [2009]. *What intelligence tests miss: the psychology of rational thought* (Yale University Press, New Haven and London).

Stanovich, K. E. and West, R.F. [1998] Individual differences in rational thought. *Journal of Experimental Psychology: General* **127**, 161-88.

Tooby, J., and Cosmides, L. [1992]. The psychological foundations of culture, in Barkow, J. H., Cosmides, L. and Tooby, J. (eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (Oxford University Press, Oxford) pp. 19-136.

Trivers, R. [2000] The Elements of a Scientific Theory of Self-Deception. *Annals of the New York Academy of Sciences* **97**, 114-131.

Uddin, L. Q., Iacoboni, M., Lange, C. and Keenan, J. P. [2007] The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences* **11**(4), 153-157.

Vinge, V. [1993] The coming technological singularity: How to survive in the post-human era. *Whole Earth Review*, Winter 1993.

Vinge, V. [2008] Signs of the Singularity, *IEEE Spectrum*, **45**(6), 76-82.

Yudkowsky, E. [2001a] Creating Friendly AI. `http://www.singinst.org/upload/CFAI/`.

Yudkowsky, E. [2001b] Staring into the Singularity. `http://sysopmind.com/singularity.html`.

Yudkowsky, E. [2007] Levels of Organization in General Intelligences, in Goertzel, B. and Pennachin, C. (eds.), *Artificial General Intelligence* (Springer-Verlag).

Yudkowsky, E. [2008a]. Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Bostrom, N. and Cirkovic, M. M. (eds.), *Global Catastrophic Risks* (Oxford University Press, Oxford), pp. 308-343.

Yudkowsky, E. [2008b] Hard Takeoff. `http://lesswrong.com/lw/wf/hard_takeoff/`.