Loosemore, R.P.W. & Goertzel, B. (2011). *Why an Intelligence Explosion is Probable.* H+ Magazine. March 7 2011. http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/

Why an Intelligence Explosion is Probable

Richard Loosemore[1] and Ben Goertzel[2],

[1] Department of Mathematical and Physical Sciences,
Wells College, Aurora NY 13026 USA,
[2] Adjunct Professor of Cognitive Science, Xiamen University, China
rloosemore@wells.edu, ben@goertzel.org

Abstract. If a future Artificial Intelligence were to reach the level of human intelligence there is a possibility that it would be able to rapidly redesign itself until its own capabilities far exceeded those of human beings. We analyze the principle factors that might govern the rapidity of this 'intelligence explosion' process. We argue that if degree of intelligence is defined using a relatively uncontroversial measure that involves only the relative speed of thought with respect to that of the human mind, and if an intelligence explosion is defined as a thousandfold increase in speed, then there are compelling reasons to believe that none of the barriers to this process look plausible, and therefore an intelligence explosion is highly likely.

Keywords: Intelligence explosion. Artificial general intelligence. Recursive self-improvement. Bottlenecks. Limitations.

1 Introduction

One of the earliest incarnations of the contemporary Singularity concept was I.J. Good's concept of the "intelligence explosion," articulated in 1965:

> *Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.*

We consider Good's vision quite plausible but, unsurprisingly, not all futurist thinkers agree. Skeptics often cite limiting factors that could stop an intelligence explosion from happening, and in a recent post on the Extropy email discussion list, the futurist Anders Sandberg articulated some of those possible limiting factors, in a particularly clear way:

> *One of the things that struck me during our Winter Intelligence workshop on intelligence explosions was how confident some people were about the speed of recursive self-improvement of AIs, brain emulation collectives or*

economies. Some thought it was going to be fast in comparison to societal adaptation and development timescales (creating a winner takes all situation), some thought it would be slow enough for multiple superintelligent agents to emerge. This issue is at the root of many key questions about the singularity (One superintelligence or many? How much does friendliness matter?).

It would be interesting to hear this list's take on it: what do you think is the key limiting factor for how fast intelligence can amplify itself?

- *Economic growth rate*
- *Investment availability*
- *Gathering of empirical information (experimentation, interacting with an environment)*
- *Software complexity*
- *Hardware demands vs. available hardware*
- *Bandwidth*
- *Lightspeed lags*

Clearly many more can be suggested. But which bottlenecks are the most limiting, and how can this be ascertained?"

We are grateful to Sandberg for presenting this list of questions because it makes it especially straightforward for us to provide a clear refutation, in this article, of the case against the viability of an intelligence explosion. We explain here why these bottlenecks (and some others commonly mentioned, such as the possible foundation of human-level intelligence in quantum mechanics) are unlikely to be significant issues, and thus why, as I.J. Good predicted, an intelligence explosion is indeed a very likely outcome.

The One Clear Prerequisite for an Intelligence Explosion

To begin, we need to delimit the scope and background assumptions of our argument. In particular, it is important to specify what kind of intelligent system would be capable of generating an intelligence explosion.

According to our interpretation, there is one absolute prerequisite for an explosion to occur, and that is that an artificial general intelligence (AGI) must become smart enough to understand its own design. In fact, by choosing to label it an "artificial general intelligence" we have already said, implicitly, that it will be capable of self understanding, since the definition of an AGI is that it has a broad set of intellectual capabilities that include all the forms of intelligence that we humans possess—and at least some humans, at that point, would be able to understand AGI design.

But even among humans there are variations in skill level and knowledge, so the AGI that triggers the explosion must have a sufficiently advanced intelligence that it can think analytically and imaginatively about how to manipulate and improve the design of intelligent systems. It is possible that not all humans are able to do this, so an AGI that met the bare minimum requirements for AGI-hood—say, a system smart enough to be a general household factotum—would not necessarily have the ability to work in an AGI research laboratory. Without an advanced AGI of the latter sort, there would

be no explosion, just growth as usual, because the rate-limiting step would still be the depth and speed at which humans can think.

The sort of fully-capable AGI we're referring to might be called a "seed AGI", but we prefer to use the less dramatic phrase "self-understanding, human-level AGI." This term, though accurate, is still rather cumbersome, so we will sometimes use the phrase "the first real AGI" or just "the first AGI" to denote the same idea. In effect, we are taking the position that for something to be a proper artificial general intelligence it has to be capable of competing with the best that the human intellect can achieve, rather than being limited to a bare minimum. So the "first AGI" would be capable of initiating an intelligence explosion.

Distinguishing the Explosion from the Preceding Build-Up

Given that the essential prerequisite for an explosion to begin would be the availability of the first self-understanding, human-level AGI, does it make sense to talk about the period leading up to that arrival—the period during which that first real AGI was being developed and trained—as part of the intelligence explosion proper? We would argue that this is not appropriate, and that the true start of the explosion period should be considered to be the moment when a sufficiently well qualified AGI turns up for work at an AGI research laboratory. This may be different from the way some others use the term, but it seems consistent with I.J. Good's original usage. So our concern here is to argue for the high probability of an intelligence explosion, given the assumption that a self-understanding, human-level AGI has been created.

By enforcing this distinction, we are trying to avoid possible confusion with the parallel (and extensive!) debate about whether a self-understanding, human-level AGI can be built at all. Questions about whether an AGI with "seed level capability" can plausibly be constructed, or how long it might take to arrive, are of course quite different. A spectrum of opinions on this issue, from a survey of AGI researchers at a 2009 AGI conference, were presented in a 2010 H+ magazine article In that survey, of an admittedly biased sample, a majority felt that an AGI with this capability could be achieved by the middle of this century, though a substantial plurality felt it was likely to happen much further out. Ray Kurzweil has also elaborated some well-known arguments in favor of the viability of AGI of this sort, based purely on extrapolating technology trends. While we have no shortage of our own thoughts and arguments on this matter, we will leave them aside for the purpose of the present paper.

It is arguable that the "intelligence explosion" as we consider it here is merely a subset of a much larger intelligence explosion that has been happening for a long time. You could redefine terms so as to say, for example, that

- **Phase 1** of the intelligence explosion occurred before the evolution of humans
- **Phase 2** occurred during the evolution of human culture

- **Phase 3 i**s Good's intelligence explosion, to occur after we have human-level AGIs

This would also be a meaningful usage of the term "intelligence explosion", but here we are taking our cue from Good's usage, and using the term "intelligence explosion" to refer to "Phase 3″ only.
While acknowledging the value of understanding the historical underpinnings of our current and future situation, we also believe the coming Good-esque "Phase 3 intelligence explosion" is a qualitatively new and different phenomenon from a *human* perspective, and hence deserves distinguished terminology and treatment.

What Constitutes an "Explosion"?

How big and how long and how fast would the explosion have to be to count as an "explosion"?

Good's original notion had more to do with the explosion's beginning than its end, or its extent, or the speed of its middle or later phases. His point was that in a short space of time a human-level AGI would probably explode into a significantly transhuman AGI, but he did not try to argue that subsequent improvements would continue without limit. We, like Good, are primarily interested in the explosion from human-level AGI to an AGI with, very loosely speaking, a level of general intelligence 2-3 orders of magnitude greater than the human level (say, 100H or 1,000H, using 1H to denote human-level general intelligence). This is not because we are necessarily skeptical of the explosion continuing beyond such a point, but rather because pursuing the notion beyond that seems a stretch of humanity's current intellectual framework.

Our reasoning, here, is that if an AGI were to increase its capacity to carry out scientific and technological research, to such a degree that it was discovering new knowledge and inventions at a rate 100 or 1,000 times the rate at which humans now do those things, we would find that kind of world unimaginably more intense than any future in which humans were doing the inventing. In a 1,000H world, AGI scientists could go from high-school knowledge of physics to the invention of relativity in a single day (assuming, for the moment, that the factor of 1,000 was all in the speed of thought—an assumption we will examine in more detail later). That kind of scenario is dramatically different from a world of purely human inventiveness—no matter how far humans might improve themselves in the future, without AGI, its seems unlikely there will ever be a time when a future Einstein would wake up one morning with a child's knowledge of science and then go on to conceive the theory of relativity by the following day—so it seems safe to call that an "intelligence explosion."

This still leaves the question of how *fast* it has to arrive, to be considered explosive. Would it be enough for the first AGI to go from 1H to 1,000H in the course of a century, or does it have to happen much quicker, to qualify?

Perhaps there is no need to rush to judgment on this point. Even a century-long climb up to the 1,000H level would mean that the world would be very different for the rest

of history. The simplest position to take, we suggest, is that if the human species can get to the point where it is creating new types of intelligence that are themselves creating intelligences of greater power, then this is something new in the world (because at the moment all we can do is create human babies of power 1H), so even if this process happened rather slowly, it would still be an explosion of sorts. It might not be a Big Bang, but it would at least be a period of Inflation, and both could eventually lead to a 1,000H world.

Defining Intelligence (Or Not)

To talk about an intelligence explosion, one has to know what one means by "intelligence" as well as by "explosion". So it's worth reflecting that there are currently no measures of general intelligence that are precise, objectively defined and broadly extensible beyond the human scope.

However, since "intelligence explosion" is a qualitative concept, we believe the commonsense qualitative understanding of intelligence suffices. We can address Sandberg's potential bottlenecks in some detail without needing a precise measure, and we believe that little is lost by avoiding the issue. We will say that an intelligence explosion is something with the potential to create AGI systems as far beyond humans as humans are beyond mice or cockroaches, but we will not try to pin down exactly how far away the mice and cockroaches really are.

Key Properties of the Intelligence Explosion

Before we get into a detailed analysis of the specific factors on Sandberg's list, some general clarifications regarding the nature of the intelligence explosion will be helpful. (Please bear with us! These are subtle matters and it's important to formulate them carefully….)

Inherent Uncertainty. Although we can try our best to understand how an intelligence explosion might happen, the truth is that there are too many interactions between the factors for any kind of reliable conclusion to be reached. This is a complex-system interaction in which even the tiniest, least-anticipated factor may turn out to be either the rate-limiting step or the spark that starts the fire. So there is an irreducible uncertainty involved here, and we should be wary of promoting conclusions that seem too firm.

General versus Special Arguments for an Intelligence Explosion. There are two ways to address the question of whether or not an intelligence explosion is likely to occur. One is based on quite general considerations. The other involves looking at specific pathways to AGI. An AGI researcher (such as either of the authors) might believe they understand a great deal of the technical work that needs to be done to create an intelligence explosion, so they may be confident of the plausibility of the idea for that reason alone. We will restrict ourselves here to the first kind of argument, which is easier to make in a relatively non-controversial way, and leave aside any factors that might arise from our own understanding about how to build an AGI.

The "Bruce Wayne" Scenario. When the first self-understanding, human-level AGI system is built, it is unlikely to be the creation of a lone inventor working in a shed at the bottom of the garden, who manages to produce the finished product without telling anyone. Very few of the "lone inventor" (or "Bruce Wayne") scenarios seem plausible. As communication technology advances and causes cultural shifts, technological progress is increasingly tied to rapid communication of information between various parties. It is unlikely that a single inventor would be able to dramatically outpace multi-person teams working on similar projects; and also unlikely that a multi-person team would successfully keep such a difficult and time-consuming project secret, given the nature of modern technology culture.

Unrecognized Invention. It also seems quite implausible that the invention of a human-level, self-understanding AGI would be followed by a period in which the invention just sits on a shelf with nobody bothering to pick it up. The AGI situation would probably not resemble the early reception of inventions like the telephone or phonograph, where the full potential of the invention was largely unrecognized. We live in an era in which practically-demonstrated technological advances are broadly and enthusiastically communicated, and receive ample investment of dollars and expertise. AGI receives relatively little funding now, for a combination of reasons, but it is implausible to expect this situation to continue in the scenario where highly technically capable human-level AGI systems exist. This pertains directly to the economic objections on Sandberg's list, as we will elaborate below.

Hardware Requirements. When the first human-level AGI is developed, it will either require a supercomputer-level of hardware resources, or it will be achievable with much less. This is an important dichotomy to consider, because world-class supercomputer hardware is not something that can quickly be duplicated on a large scale. We could make perhaps hundreds of such machines, with a massive effort, but probably not a million of them in a couple of years.

Smarter versus Faster. There are two possible types of intelligence speedup: one due to faster operation of an intelligent system (clock speed increase) and one due to an improvement in the type of mechanisms that implement the thought processes ("depth of thought" increase). Obviously both could occur at once (and there may be significant synergies), but the latter is ostensibly more difficult to achieve, and may be subject to fundamental limits that we do not understand. Speeding up the hardware, on the other hand, is something that has been going on for a long time and is more mundane and reliable. Notice that both routes lead to greater "intelligence," because even a human level of thinking and creativity would be more effective if it were happening a thousand times faster than it does now.

It seems quite possible that the general class of AGI systems can be architected to take better advantage of improved hardware than would be the case with intelligent systems very narrowly imitative of the human brain. But even if this is not the case, brute hardware speedup can still yield dramatic intelligent improvement.

Public Perception. The way an intelligence explosion presents itself to human society will depend strongly on the rate of the explosion in the period shortly after the

development of the first self-understanding human-level AGI. For instance, if the first such AGI takes five years to "double" its intelligence, this is a very different matter than if it takes two months. A five-year time frame could easily arise, for example, if the first AGI required an extremely expensive supercomputer based on unusual hardware, and the owners of this hardware were to move slowly. On the other hand, a two-month time frame could more easily arise if the initial AGI were created using open source software and commodity hardware, so that a doubling of intelligence only required addition of more hardware and a modest number of software changes. In the former case, there would be more time for governments, corporations and individuals to adapt to the reality of the intelligence explosion before it reached dramatically transhuman levels of intelligence. In the latter case, the intelligence explosion would strike the human race more suddenly. But this potentially large difference in human perception of the events would correspond to a fairly minor difference in terms of the underlying processes driving the intelligence explosion.

So – now, finally, with all the preliminaries behind us, we will move on to deal with the specific factors on Sandberg's list, one by one, explaining in simple terms why each is not actually likely to be a significant bottleneck. There is much more that could be said about each of these, but our aim here is to lay out the main points in a compact way.

Objection 1: Economic Growth Rate and Investment Availability

The arrival, or imminent arrival, of human-level, self-understanding AGI systems would clearly have dramatic implications for the world economy. It seems inevitable that these dramatic implications would be sufficient to offset any factors related to the economic growth rate at the time that AGI began to appear. Assuming the continued existence of technologically advanced nations with operational technology R&D sectors, if self-understanding human-level AGI is created, then it will almost surely receive significant investment. Japan's economic growth rate, for example, is at the present time somewhat stagnant, but there can be no doubt that if any kind of powerful AGI were demonstrated, significant Japanese government and corporate funding would be put into its further development.

And even if it were not for the normal economic pressure to exploit the technology, international competitiveness would undoubtedly play a strong role. If a working AGI prototype were to approach the level at which an explosion seemed possible, governments around the world would recognize that this was a critically important technology, and no effort would be spared to produce the first fully-functional AGI "before the other side does." Entire national economies might well be sublimated to the goal of developing the first superintelligent machine, in the manner of Project Apollo in the 1960s. Far from influencing the intelligence explosion, economic growth rate would be *defined* by the various AGI projects taking place around the world.

Furthermore, it seems likely that once a human-level AGI has been achieved, it will have a substantial—and immediate—practical impact on multiple industries. If an

AGI could understand its own design, it could also understand and improve other computer software, and so have a revolutionary impact on the software industry. Since the majority of financial trading on the US markets is now driven by program trading systems, it is likely that such AGI technology would rapidly become indispensible to the finance industry (typically an early adopter of any software or AI innovations). Military and espionage establishments would very likely also find a host of practical applications for such technology. So, following the achievement of self-understanding, human-level AGI, and complementing the allocation of substantial research funding aimed at outpacing the competition in achieving ever-smarter AGI, there is a great likelihood of funding aimed at practical AGI applications, which would indirectly drive core AGI research along.

The details of how this development frenzy would play out are open to debate, but we can at least be sure that the economic growth rate and investment climate in the AGI development period would quickly become irrelevant.

However, there is one interesting question left open by these considerations. At the time of writing, AGI investment around the world is noticeably weak, compared with other classes of scientific and technological investment. Is it possible that this situation will continue indefinitely, causing so little progress to be made that no viable prototype systems are built, and no investors ever believe that a real AGI is feasible?

This is hard to gauge, but as AGI researchers ourselves, our (clearly biased) opinion is that a "permanent winter" scenario is too unstable to be believable. Because of premature claims made by AI researchers in the past, a barrier to investment clearly exists in the minds of today's investors and funding agencies, but the climate already seems to be changing. And even if this apparent thaw turns out to be illusory, we still find it hard to believe that there will not eventually be an AGI investment episode comparable to the one that kicked the internet into high gear in the late 1990s. Furthermore, due to technology advanced in allied fields (computer science, programming language, simulation environments, robotics, computer hardware, neuroscience, cognitive psychology, etc.), the amount of effort required to implement advanced AGI designs is steadily decreasing – so that as time goes on, the amount of investment required to get AGI to the explosion-enabling level will keep growing less and less.

Objection 2: Inherent Slowness of Experiments and Environmental Interaction

This possible limiting factor stems from the fact that any AGI capable of starting the intelligence explosion would need to do some experimentation and interaction with the environment in order to improve itself. For example, if it wanted to reimplement itself on faster hardware (most probably the quickest route to an intelligence increase) it would have to set up a hardware research laboratory and gather new scientific data by doing experiments, some of which might proceed slowly due to limitations of experimental technology.

The key question here is this: how much of the research can be sped up by throwing large amounts of intelligence at it? This is closely related to the problem of

parallelizing a process (which is to say: you cannot make a baby nine times quicker by asking nine women to be pregnant for one month). Certain algorithmic problems are not easily solved more rapidly simply by adding more processing power, and in much the same way there might be certain crucial physical experiments that cannot be hastened by doing a parallel set of shorter experiments.

This is not a factor that we can understand fully ahead of time, because some experiments that look as though they require fundamentally slow physical processes—like waiting for a silicon crystal to grow, so we can study a chip fabrication mechanism—may actually be dependent on the intelligence of the experimenter, in ways that we cannot anticipate. It could be that instead of waiting for the chips to grow at their own speed, the AGI could do some clever micro-experiments that yield the same information faster.

The increasing amount of work being done on nanoscale engineering would seem to reinforce this point—many processes that are relatively slow today could be done radically faster using nanoscale solutions. And it is certainly feasible that advanced AGI could accelerate nanotechnology research, thus initiating a "virtuous cycle" where AGI and nanotech research respectively push each other forward (as foreseen by nanotech pioneer Josh Hall). As current physics theory does not even rule out more outlandish possibilities like femtotechnology, it certainly does not suggest the existence of absolute physical limits on experimentation speed existing anywhere near the realm of contemporary science.

Clearly, there is significant uncertainty in regards to this aspect of future AGI development. One observation, however, seems to cut through much of the uncertainty. Of all the ingredients that determine how fast empirical scientific research can be carried out, we know that in today's world the intelligence and thinking speed of the scientists themselves must be one of the most important. Anyone involved with science and technology R&D would probably agree that in our present state of technological sophistication, advanced research projects are strongly limited by the availability and cost of intelligent and experienced scientists.

But if research labs around the world have stopped throwing more scientists at problems they want to solve, because the latter are unobtainable or too expensive, would it be likely that those research labs are also, quite independently, at the limit for the physical rate at which experiments can be carried out? It seems hard to believe that both of these limits would have been reached at the same time, because they do not seem to be independently optimizable. If the two factors of experiment speed and scientist availability could be independantly optimized, this would mean that even in a situation where there was a shortage of scientists, we could still be sure that we had discovered all of the fastest possible experimental techniques, with no room for inventing new, ingenious techniques that get over the physical-experiment-speed limits. In fact, however, we have every reason to believe that if we were to double the number of scientists on the planet at the moment, some of them would discover new ways to conduct experiments, exceeding some of the current speed limits. If that were not true, it would mean that we had quite coincidentally reached the limits of

science talent and physical speed of data collecting at the same time—a coincidence that we do not find plausible.

This picture of the current situation seems consistent with anecdotal reports: companies complain that research staff are expensive and in short supply; they do not complain that nature is just too slow. It seems generally accepted, in practice, that with the addition of more researchers to an area of inquiry, methods of speeding up and otherwise improving processes can be found.

So based on the actual practice of science and engineering today (as well as known physical theory), it seems most likely that any experiment-speed limits lie further up the road, out of sight. We have not reached them yet, and we lack any solid basis for speculation about exactly where they might be.

Overall, it seems we do not have concrete reasons to believe that this will be a fundamental limit that stops the intelligence explosion from taking an AGI from H (human-level general intelligence) to (say) 1,000 H. Increases in speed within that range (for computer hardware, for example) are already expected, even without large numbers of AGI systems helping out, so it would seem that physical limits, by themselves, would be very unlikely to stop an explosion from 1H to 1,000 H.

Objection 3: Software Complexity

This factor is about the complexity of the software that an AGI must develop in order to explode its intelligence. The premise behind this supposed bottleneck is that even an AGI with self-knowledge finds it hard to cope with the fabulous complexity of the problem of improving its own software.

This seems implausible as a limiting factor, because the AGI could always leave the software alone and develop faster hardware. So long as the AGI can find a substrate that gives it a thousand-fold increase in clock speed, we have the possibility for a significant intelligence explosion.

Arguing that software complexity will stop the *first* self-understanding, human-level AGI from being built is a different matter. It may stop an intelligence explosion from happening by stopping the precursor events, but we take that to be a different type of question. As we explained earlier, one premise of the present analysis is that an AGI can actually be built. It would take more space than is available here to properly address that question.

It furthermore seems likely that, if an AGI system is able to comprehend its own software as well as a human being can, it will be able to improve that software significantly beyond what humans have been able to do. This is because in many ways, digital computer infrastructure is more suitable to software development than the human brain's wetware. And AGI software may be able to interface directly with programming language interpreters, formal verification systems and other programming-related software, in ways that the human brain cannot. In that way the software complexity issues faced by human programmers would be significantly

mitigated for human-level AGI systems. However, this is not a 100% critical point for our arguments, because even if software complexity remains a severe difficulty for a self-understanding, human-level AGI system, we can always fall back to arguments based on clock speed.

Objection 4: Hardware Requirements

We have already mentioned that much depends on whether the first AGI requires a large, world-class supercomputer, or whether it can be done on something much smaller.

This is something that could limit the initial speed of the explosion, because one of the critical factors would be the number of copies of the first AGI that can be created. Why would this be critical? Because the ability to *copy* the intelligence of a fully developed, experienced AGI is one of the most significant mechanisms at the core of an intelligence explosion. We cannot do this copying of adult, skilled humans, so human geniuses have to be rebuilt from scratch every generation. But if one AGI were to learn to be a world expert in some important field, it could be cloned any number of times to yield an instant community of collaborating experts.

However, if the first AGI had to be implemented on a supercomputer, that would make it hard to replicate the AGI on a huge scale, and the intelligence explosion would be slowed down because the replication rate would play a strong role in determining the intelligence-production rate.

However, as time went on, the rate of replication would grow, as hardware costs declined. This would mean that the rate of arrival of high-grade intelligence would increase in the years following the start of this process. That intelligence would then be used to improve the design of the AGIs (at the very least, increasing the rate of new-and-faster-hardware production), which would have a positive feedback effect on the intelligence production rate.

So if there was a supercomputer-hardware requirement for the first AGI, we would see this as something that would only dampen the initial stages of the explosion. Positive feedback after that would eventually lead to an explosion anyway.

If, on the other hand, the initial hardware requirements turn out to be modest (as they could very well be), the explosion would come out of the gate at full speed.

Objection 5: Bandwidth

In addition to the aforementioned cloning of adult AGIs, which would allow the multiplication of knowledge in ways not currently available in humans, there is also the fact that AGIs could communicate with one another using high-bandwidth channels. This is *inter-AGI bandwidth*, and it is one of the two types of bandwidth factors that could affect the intelligence explosion.

Quite apart from the communication speed between AGI systems, there might also be bandwidth limits inside a single AGI, which could make it difficult to augment the intelligence of a single system. This is *intra-AGI bandwidth*.

The first one—inter-AGI bandwidth—is unlikely to have a strong impact on an intelligence explosion because there are so many research issues that can be split into separably-addressible components. Bandwidth between the AGIs would only become apparent if we started to notice AGIs sitting around with no work to do on the intelligence amplification project, because they had reached an unavoidable stopping point and were waiting for other AGIs to get a free channel to talk to them. Given the number of different aspects of intelligence and computation that could be improved, this idea seems profoundly unlikely.

Intra-AGI bandwidth is another matter. One example of a situation in which internal bandwidth could be a limiting factor would be if the AGI's working memory capacity were dependent on the need for total connectivity—everything connected to everything else—in a critical component of the system. If this case, we might find that we could not boost working memory very much in an AGI because the bandwidth requirements would increase explosively. This kind of restriction on the design of working memory might have a significant effect on the system's depth of thought.

However, notice that such factors may not inhibit the initial phase of an explosion, because the clock speed, not the depth of thought, of the AGI may be improvable by several orders of magnitude before bandwidth limits kick in. The main element of the reasoning behind this is the observation that neural signal speed is so slow. If a brain-like AGI system (not necessarily a whole brain emulation, but just something that replicated the high-level functionality of the brain) could be built using components that kept the same type of processing demands, and the same signal speed as neurons, then we would be looking at a human-level AGI in which information packets were being exchanged once every millisecond. In that kind of system there would then be plenty of room to develop faster signal speeds and increase the intelligence of the system. The processing elements would also have to go faster, if they were not idling, but the point is that the bandwidth would not be the critical problem.

Objection 6: Lightspeed Lags

Here we need to consider the limits imposed by special relativity on the speed of information transmission in the physical universe. However, its implications in the context of AGI are not much different than those of bandwidth limits.

Lightspeed lags could be a significant problem if the components of the machine were physically so far apart that massive amounts of data (by assumption) were delivered with a significant delay. But they seem unlikely to be a problem in the initial few orders of magnitude of the explosion. Again, this argument derives from what we know about the brain. We know that the brain's hardware was chosen due to biochemical constraints. We are carbon-based, not silicon-and-copper-based, so there are no electronic chips in the head, only pipes filled with fluid and slow molecular

gates in the walls of the pipes. But if nature was forced to use the pipes-and-ion-channels approach, that leaves us with plenty of scope for speeding things up using silicon and copper (and this is quite apart from all the other more exotic computing substrates that are now on the horizon). If we were simply to make a transition membrane depolarization waves to silicon and copper, and if this produced a 1,000x speedup (a conservative estimate, given the intrinsic difference between the two forms of signalling), this would be an explosion worthy of the name.

The main circumstance under which this reasoning would break down would be if, for some reason, the brain is limited on two fronts simultaneously: both by the carbon implementation and by the fact that other implementations of the same basic design are limited by disruptive light-speed delays. This would mean that all non-carbon-implementations of the brain take us up close to the lightspeed limit before we get much of a speedup over the brain. This would require a coincidence of limiting factors (two limiting factors just happening to kick in at exactly the same level), that we find quite implausible, because it would imply a rather bizarre situation in which evolution tried both the biological neuron design, and a silicon implementation of the same design, and after doing a side-by-side comparison of performance, chose the one that pushed the efficiency of all the information transmission mechanisms up to their end stops.

Objection 7: Human-Level Intelligence May Require Quantum (or more exotic) Computing

Finally we consider an objection not on Sandberg's list, but raised from time to time in the popular and even scientific literature. The working assumption of the vast majority of the contemporary AGI field is that human-level intelligence can eventually be implemented on digital computers, but the laws of physics as currently understood imply that, to simulate certain physical systems without dramatic slowdown, requires special physical systems called "quantum computers" rather than ordinary digital computers.

There is currently no evidence that the human brain is a system of this nature. Of course the brain has quantum mechanics at its underpinnings, but there is no evidence that it displays quantum coherence at the levels directly relevant to human intelligent behavior. In fact our current understanding of physics implies that this is unlikely, since quantum coherence has not yet been observed in any similarly large and "wet" system. Furthermore, even if the human brain were shown to rely to some extent on quantum computing, this wouldn't imply that quantum computing is necessary for human-level intelligence — there are often many different ways to solve the same algorithmic problem. And (the killer counterargument), even if quantum computing *were* necessary for human-level general intelligence, that would merely delay the intelligence explosion a little, while suitable quantum computing hardware was developed. Already the development of such hardware is the subject of intensive R&D.

Roger Penrose, Stuart Hameroff and a few others have argued that human intelligence may even rely on some form of "quantum gravity computing", going beyond what

ordinary quantum computing is capable of. But this is really a complete blue-sky speculation with no foundation in current science, so not worth discussing in detail in this context. The simpler versions of this claim may be treated according to the same arguments as we've presented above regarding quantum computing. The strongest versions of the claim include an argument that human-level intelligence relies on extremely powerful mathematical notions of "hyper-Turing computation" exceeding the scope of current (or maybe any possible) physics theories; but here we verge on mysticism, since it's arguable that no set of scientific data could ever validate or refute such an hypothesis.

The Path from AGI to Intelligence Explosion Seems Clear

Summing up, then — the conclusion of our relatively detailed analysis of Sandberg's objections is that there is currently no good reason to believe that once a human-level AGI capable of understanding its own design is achieved, an intelligence explosion will fail to ensue.

The operative definition of "intelligence explosion" that we have assumed here involves an increase of the speed of thought (and perhaps also the "depth of thought") of about two or three orders of magnitude. If someone were to insist that a real intelligence explosion had to involve million-fold or trillion-fold increases in intelligence, we think that no amount of analysis, at this stage, could yield sensible conclusions. But since an AGI with intelligence = 1000 H might well cause the next thousand years of new science and technology to arrive in one year (assuming that the speed of physical experimentation did not become a significant factor within that range), it would be churlish, we suggest, not to call that an "explosion". An intelligence explosion of such magnitude would bring us into a domain that our current science, technology and conceptual framework are not equipped to deal with; so prediction beyond this stage is best done once the intelligence explosion has already progressed significantly.

Of course, even if the above analysis is correct, there is a great deal we do not understand about the intelligence explosion, and many of these particulars will remain opaque until we know precisely what sort of AGI system will launch the explosion. But our view is that the likelihood of transition from a self-understanding human-level AGI to an intelligence explosion should not presently be a subject of serious doubt. And we also feel that the creation of a self-understanding human-level AGI is a high-probability outcome, though this is a more commonplace assertion and we have not sought to repeat the arguments in its favor here.

Of course, if our analysis is correct, there are all sorts of dramatic implications for science, society and humanity (and beyond) — but many of these have been discussed elsewhere, and reviewing this body of thought is not our purpose here. These implications are worth deeply considering — but the first thing is to very clearly understand ***that the intelligence explosion is very probably coming***, just as I.J. Good foresaw.