Is Brain Emulation Dangerous?

Peter Eckersley and Anders Sandberg

Brain emulation is a hypothetical but extremely transformative technology which has a non-zero chance of appearing during the next century. This paper investigates whether such a technology would have any predictable characteristics that make it catastrophically dangerous, and whether there are any policy levers which might be used to make it safer.

We conclude that the riskiness of brain emulation probably depends on the order of the preceding research trajectory. Broadly speaking, it appears safer for brain emulation to happen sooner, because slower CPUs would make the technology's impact more gradual. It may also be safer if brains are scanned before they are fully understood from a neuroscience perspective, thereby increasing the initial population of emulations, although this prediction is weaker and more scenario-dependent.

The risks posed by brain emulation also seem strongly connected to questions about the balance of power between attackers and defenders in computer security contests. If economic property rights in CPU cycles are essentially enforceable, emulation appears to be comparatively safe; if CPU cycles are ultimately easy to steal, the appearance of brain emulation is more likely to be a destabilising development for human geopolitics.

Furthermore, if the computers used to run emulations can be kept secure, then it appears that making brain emulation technologies "open" would make them safer. If, however, computer insecurity is deep and unavoidable, openness may actually be more dangerous. We point to some arguments that suggest the former may be true, tentatively implying that it would be good policy to work towards brain emulation using open scientific methodology and free/open source software codebases.

Why Study Artificial Intelligence Risks before Artificial Intelligence is possible?

The proposition that Artificial General Intelligence (AGI) might pose a catastrophic or existential threat to life on earth sounds more like a plotline from science fiction than a serious object of academic study. Be that as it may, actuarial risk assessment tells us that even if we assign numerically small probabilities to such events, they could remain serious enough to deserve study and mitigation. Un-intuitively, we should worry about AGI catastrophes even if we don't think that they are the likely course of events. The study of these scenarios should be thought of as an insurance policy against an unlikely but serious adverse event. Existential risks may in fact be strictly more important than any other global public good [Nick Bostrom. Existential Risk Prevention as Global Priority. Global Policy Volume 4, Issue 1, February 2013]. A small

but growing literature has started that project [See for example Yudkowsky, Eliezer. "Artificial Intelligence as a positive and negative factor in global risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic, 308-345. Oxford University Press, 2008. Omohundro, S. "The basic AI drives." Edited by B. G. P. Wang. *Proceedings of the First AGI Conference* (Frontiers in Artificial Intelligence and Applications, IOS Press.) 171 (2008). Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. Berlin: Springer. Bostrom, Nick. *Superintelligence: an analysis of the coming machine intelligence revolution*. in preparation, 2013.].

One objection to this line of reasoning is that it is too soon for us to begin. As of 2012, there are no research projects that can credibly claim to be close to producing an AGI, so it would be necessary to make meaningful predictions about a phenomenon whose details will be unknown until the medium- to long-term future. This far out, it is extremely difficult to reason accurately about the actions and motivations of AGIs. In particular, the diversity of possibilities is astonishingly large. Differences in the design, education, early experiences and social surroundings of conceivable AGIs create a space of possible intelligences and personalities far larger than that of human intelligences and personalities, which are in part constrained by our biology.

Given this profound difficulty in saying much about what AGIs would be like produces a corresponding difficulty in evaluating any catastrophic or existential risks that their appearance might induce, and if we cannot evaluate risks we have little hope of mitigating them sensibly.

This paper will avoid that vast degree of unpredictability by focusing on one possible subtype of artificial general intelligence: human brains that have been emulated by computers. As a possible future technology, emulations of human brains are slightly more predictable than systems which are built from scratch, since they would at least at first be a combination of things we know quite a lot about: human personalities running inside computers.

A simple taxonomy of Artificial General Intelligence

There are at least three different ways that artificial intelligence research might succeed in building an Artificial General Intelligence (AGI) with capabilities for learning, problem-solving and intellectual labour comparable to those of humans:

1. "Designed" AGI
2. Evolved AGI
3. Whole Brain Emulation

The first two categories would constitute success by different strands of traditional artificial intelligence research: either designing a system with sufficient cleverness, complexity and flexibility that it demonstrates intelligence, or building a framework for some very abstract

algorithm (such as a neural network or an evolutionary program) to find the ingredients of intelligence by trial, error and combination.

The third kind of success, Whole Brain Emulation, is relatively new as a serious research objective. This project would involve taking an individual human's brain, scanning its entire neural (and perhaps neurochemical) structure into a computer, and running an algorithm to emulate that brain's behaviour, using virtual reality systems and an emulated body for sensory input and output.

The Capabilities of Emulated Humans

It is not the intention of this article to discuss whether the emulation of human intelligence is possible, feasible, or likely in any given timeframe. That matter is taken up at length by Sandberg and Bostrom (2008). [Sandberg, A. & Bostrom, N. (2008): Whole Brain Emulation: A Roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University, http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf.] See also [David J. Chalmers. The Singularity: A Philosophical Analysis. The Journal of Consciousness Studies 17:7-65, 2010. Anders Sandberg, Feasibility of Whole Brain Emulation, in Philosophy and Theory of Artificial Intelligence, ed. Vincent Müller, SAPERE 5, pp. 251--264.] It suffices to note here that there is a significant probability that such emulations are possible, and that the implications may be large enough to be of interest regardless of whether the probability that emulations occur in the next century is 1% or 99%.

We can predict certain capabilities of emulations of human beings and the world they will exist in with high probability. The main precondition for these predictions is that sufficient computational resources are available to run a number of these emulations simultaneously. There are many other capabilities which emulations might develop, but these are the least speculative.

Emulations can be copied

If human thought processes can be correctly emulated by computers, then their internal states can be represented as digital data. Digital data can be copied. It follows that an emulated entity can at any time be copied, and if a different set of computers begins emulating the copy, two independent versions of the original entity can arise.

From an economical perspective this property is very important: it allows human capital to be multiplied easily, rather than relying on slow human reproduction followed by expensive (and slow) education. [Hanson, R. (2008). Economics of the singularity. *IEEE Spectrum,* 37-42.]

Copying also makes it possible to keep backups. Given sufficient resources, this property might often make death a local phenomenon rather than a global one: a loss of some post-backup experience, rather than the end of the agent’s existence.

Some computer architectures may by design or accident make their internal data harder to extract, for example by storing the emulation as a wiring pattern of FPGAs or a trusted computing system disallowing copying. Whether such architectures would become dominant depends on whether they would afford efficiency or perceived security advantages outweighing the cumbersomeness (for developing and debugging early emulations there is an obvious need for direct access and possibly copying, and such needs may remain long after the technology is mature). However, there would still exist a "seed" dataset that is used to initialize the restricted hardware, and unless that was stored within an equally locked structure it would be copyable.

There is also a practical bandwidth limitation: the size of emulations is likely to be on the order of tens of terabytes or larger [Sandberg and Bostrom (2008)], making online distribution slow (compared to computer speeds) at least for the next few decades. Local copying within the same data center might be far more efficient than copying "abroad", producing a localization tendency.

Whether emulations *want* to be copied is an individual matter. No doubt the people willing to undergo scanning are more likely than the average person to have philosophical views compatible with functionalism and multiple realizability, but whether that means they will desire many copies is still highly individual. However, as noted by Robin Hanson and others, it is enough that *some* emulations are fine with copying for there to be many copies.

Emulations can be erased by network attacks

Similar to the copying issue, existing as digital data makes it possible to modify or erase brain emulations rapidly without changing the underlying hardware. Emulations can be instantly deleted by whoever controls the hardware or the operating system, including the distant author of a virus or other type of malware. It may be possible for resourceful emulations to build defenses against such attacks, such as offline backups managed by humans or air-gapped copies of themselves, or it may be that attackers will typically have ways to defeat these protections.

Humans are of course similarly vulnerable to assassination, although it is rare for this threat to exist with the same level of distance and anonymity that malware authors commonly attain. It is also relevant that an emulation which erases another emulation may be able to take those CPU cycles for itself.

Emulations will probably be fast

The task of emulating neurons in a brain is highly parallelizable. In simple terms, CPU A can be busy emulating one region of a brain, while CPU B can be emulating regions in another portion of the brain. As more CPUs are made available, the number of neurons each CPU is responsible for decreases, thereby allowing the emulation to run faster.

We do not know what the limit of this "speedup" process is. It is possible that silicon CPUs are incapable of emulating a brain in real-time. It is more likely that brain emulation in faster-than-real time is possible. The principal reason for thinking that is the characteristic timescale on which neurons appear to communicate, which is on the order of 100Hz or slower [Steriade M, Timofeev I, Durmuller N, Grenier F (1998) Dynamic properties of corticothalamic neurons and local circuit interneurons generating fast rhythmic (30–40 Hz) spike bursts. *J Neurophysiol* 79:483–490.], with conduction delays between a few milliseconds and up to a hundred milliseconds ["Axonal conduction delays" by Harvey A. Swadlow and Stephen G. Waxman, *Scholarpedia*, 7(6):1451. 2012 http://www.scholarpedia.org/article/Axonal_conduction_delays]. Digital signals can travel very long distances during each of those cycles, meaning that a very large number of CPU cores can be simultaneously brought to bear on a single emulation task. Modern CPUs are also many orders of magnitude faster (with gigahertz speeds rather than hektoherz speeds), allowing the same system to be simulated a higher rate than in nature.

A mechanism for communicating between the CPUs A and B that were emulating interdependent neurons A' and B' with a latency of *k* Hz should in principle be able to support a speed up of between *k*/200 and *k*/100 times, depending on the complexity of the circuit [footnote: *k/*200 case is where a full "cycle" is necessary to compute the consequences of a neural input; *k*/100 is where the computation is trivial, or all 2 or more computations can be performed in parallel, while waiting to know which of them is valid].

Existing digital systems can comfortably support signal propagation at around one-tenth the speed of light, though there are some technologies that are faster [Richard T. Chang *et al.*, 2003 Near Speed-of-Light Singaling Over On-Chip Electrical Interconnects, *IEEE Journal of Solid-State Circuits,* Vol 38 No 5 http://www.ece.ucsb.edu/yuegroup/Publications/JSSC03_1.pdf]. It follows that the distance between CPU cores working to emulate the neighbouring neurons A' and B' would need to be less than *d* =(3x10^8 * 0.1) / 200 = 1.5 x 10^5 m apart. This strongly suggests that the main bound on the speed of first-generation brain emulations would be the number of CPUs available for the task, or the amount of money available to purchase them. The bound might be high or low, relative to humans, but it would be proportional to the availability of these resources. Over time, these bounds are likely to increase according to Moore's law and a larger installed computer base.

Dedicated hardware particularly suited for brain emulation might also provide speedups over generic CPUs, especially if brain emulation is a profitable enough business to motivate large servers. Typically such specializations provide between one and two orders of magnitude speedup. [Sandberg and Bostrom (2008)]

Emulation autonomy would be fragile

Suppose an agent Alice (who might be human, or an emulation) possesses a digital copy of the full neural state of an emulation, who we will call Aesop. Suppose further that Alice has access to enough storage and computational resources to make further copies of the emulation and run

some of these copies.

Alice can instantiate Aesop. She can control the virtual reality environment in which Aesop finds himself: his senses (or his attempts to use communications systems) can only tell him about the world to the extent that Alice allows this. Furthermore, she could construct fake stories and details of reality to misdirect him. If necessary, she could slow or freeze the rate at which he is emulated, in order to determine off-line the most convincing virtual reality response to one of his actions.

It seems that Alice can persuade Aesop to do almost anything. In particular, she can copy a state, and then attempt to persuade him in way A. If he refuses, she can restore the old state, and then attempt to use persuasive method B. There is no bound on the number of persuasive techniques she might try.

The instant that Alice has persuaded Aesop to perform a single task for her, she can pause and make a copy of his mental state before she tells him the details of the task. Thereafter, she can reinstantiate that state and hand Aesop a different problem to solve.

The best Aesop could do to defend himself against Alice's predations would be to constantly insist on interacting with the physical Earth in complicated ways, hoping that Alice could not fake such interactions. But he would be constantly vulnerable to trickery, constantly in danger of performing tasks that served Alice's ends rather than his own.

Once Alice has done this, Aesop appears to be virtually enslaved to her. Aesop, or at least this copy of him, no longer possesses autonomy.

An emulation that owns or has effective control over the hardware necessary for her own existence would normally enjoy autonomy. But any time that the physical or software security of those systems was compromised, the agent would face the risk that someone might make non-autonomous, enslaveable copies of their mental states.

Presuming that the emulation was able to reassert control over her own hardware, she might now find that a copy of herself had been enslaved by someone else.

It isn't clear what views agents would have about such situations, but it is very likely that they would often be quite paranoid about such events.

Of course, one can make predictive arguments that enslaving emulations might be more difficult than presented above. Perhaps it is *not* actually practical for Alice to fake the world that Aesop thinks he is interacting with (it is too time-consuming to construct versions of data and events that Aesop would find plausible and chronologically consistent), and Aesop can at minimum tell that he has been enslaved.

A further counter-counter argument is that Alice could possibly learn to edit Aesop's internal mental states in order to make enslavement easier. With access to the emulated neural substrate Alice is able to stimulate pleasure, pain and emotion arbitrarily, monitor motivational states or attempts at deception with great precision, and perform numerous neuropsychological

tricks.

For now the question of whether emulations have fragile or not-so-fragile autonomy may be unknowable; however we believe the fragile autonomy hypothesis is somewhat more likely, and that it is also probably the more conservative as a proposition to build upon.

What Can Go Wrong With Emulated Humans

Obviously there might be many scenarios of accidental or deliberate mistreatment or misuse of emulated humans, but this paper will focus on the cases where the result is a global threat rather than a local or moral one. See [Anders Sandberg, Ethics of brain emulations, *Journal of Experimental & Theoretical Artificial Intelligence*, special issue edited by Vincent C. Müller, forthcoming 2013] for an analysis of some of the ethical issues of brain emulation.

Dynamics Leading to Existential or Global Catastrophic Risks

A Cause of Conflict

One reason that brain emulation might be dangerous is that it might increase the likelihood or severity of geopolitical strife or war.

The consensus view among those who have studied the history of war is that there is no single agreed major cause for violent conflict. [Jack S. Levy, William R. Thompson. Causes of War. John Wiley & Sons, 2009. Matthew White, *Atrocitology: Humanity's 100 Deadliest Achievements*. Cannongate Books, 2012] Unfortunately, brain emulation technologies are capable of providing many of the kinds of ingredients that are commonly regarded as contributing to the risk of war [Macartan Humphreys. Economics and Violent Conflict. Working paper UNICEF 2002. Stephen Van Evera. Causes of War: Power and the Roots of Conflict. Cornell University Press, 2013], including:

- increasing inequality (between emulations, humans who can afford and want to "become" emulations, and humans who cannot);
- groups that become marginalized (humans who cannot compete with emulations, emulations or groups of emulations that are at a disadvantage compared to other emulations);
- disruption of existing social power relationships and the creation of opportunities to establish new kinds of power;
- potential first strike-advantages and cumulative resource advantages (holding more resources increases the resource-gathering efficiency);

- the appearance of groups of intelligent beings who may empathise with each other even less than humans historically have done;
- the appearance of groups of beings with strong internal loyalty and greater willingness to 'die' for what they value; [Carl Shulman, Whole Brain Emulation and the Evolution of Superorganisms, report of The Singularity Institute, San Francisco, CA. 2010 intelligence.org/files/WBE-Superorgs.pdf]
- particularly strong triggers for racist and xenophobic prejudices;
- particularly strong triggers for vigorous religious objections;
- the creation of situations in which the scope of human rights and property rights are poorly defined and subject to dispute (and surprise).

Of course, the fact that emulations may cause these factors does not guarantee that the factors will lead to geopolitical violence. The emergence of the Internet, for instance, disrupted social power relationships and in some settings increased inequality, but does not seem to have triggered violent conflict globally. On the other hand, brain emulation might produce more of these ingredients for conflict, and much more quickly, than the Internet did.

For example, the emergence of emulation technology might make a few countries sites of unprecedented economic growth (in Hanson's model the GDP would roughly follow Moore's law with a doubling time of 18 months). Other countries, lacking the necessary computing infrastructure, legal protections, or economic desirability, would be at a significant and rapidly worsening disadvantage. Technology diffusion would be slow compared to the growth of leading edge advantage. The prudent response would be to rush to transfer the technology, but there is a risk that some would respond by attempting to disrupt or destabilise the leading countries. The leading countries would need to respond with restraint in order to avoid escalation.

Conflict can also occur as a side effect of an arms race between several countries pursuing emulation, each recognizing its game-changing potential and unwilling to let the other side gain the advantage first (or share it). Here the negative effect is not due to emulation technology itself, it merely amplifies an existing geopolitical risk by its apparent extreme potential.

While war might be the most "traditional" risk on this list, it is a very significant risk. Although the frequency of large conflicts may have declined historically [Steven Pinker, The Better Angels of Our Nature: Why Violence Has Declined" (Viking Books, 2011)], the statistical distribution of conflict fatalities show a heavy tail: the largest conflicts to date dominate the total number of fatalities.[Lewis F. Richardson. Variation of the Frequency of Fatal Quarrels With Magnitude. Journal of the American Statistical Association , Vol. 43, No. 244 (Dec., 1948), pp. 523-546. Lars-Erik Cederman. Modeling the Size of Wars: From Billiard Balls to Sandpiles. The American Political Science Review , Vol. 97, No. 1 (Feb., 2003), pp. 135-150] There is also no reason to believe that there exist a natural cut-off in conflict size (besides close to the entire world population), given current global intervention capabilities and weapons of mass destruction. This means that damage in future conflicts might well dwarf the sum total of all past conflicts.

Conflicts in a world with emulation technology are worsened by the different vulnerabilities and needs of biological and emulated humans. Biological humans are vulnerable to threats such as biological warfare that are irrelevant to emulated humans, while emulations could be attacked

using software means. Emulations are dependent on a computational infrastructure that might be easily disrupted, while traditional humans are dependent on the biosphere and the food infrastructure. These infrastructures are entwined and partially overlap, but there is enough separation that certain attacks might preferentially harm one group over the other and hence appear tempting to overconfident leaders.

In a conflict situation, the fast timescales of emulations can be destabilizing (similar to concerns about the effect of missile weaponry [Michael D. Wallace, Brian L. Crissey, Linn I. Sennott, Accidental Nuclear War: A Risk Assessment. *Journal of Peace Research* March 1986 23: 9-27, doi:10.1177/002234338602300102]). There is not much time for biological decisionmakers to negotiate with each other, change decisions or even make high quality decisions if the conflict occurs on emulation timescales.

A runaway emulation

One possible behaviour of an emulated human — and especially an emulated human with its own agency — is to seek out new computational resources as rapidly as possible, and to use those resources to accelerate or reproduce itself. Given the bounds we derive above (in section “emulations will probably be fast”), there is a chance that such emulations could benefit from very rapid positive feedback loops.

Such a scenario poses risks of at least two inherent kinds: antisocial or violent conduct by the runaway emulation, or undesirable and persistent influence of the earth's future as a result of the personality and power of that first runaway agent. For example, the emulation might exploit its speed and relatively cheap virtual existence to underbid other agents for work, becoming numerous and influential - to the point where most of the world economy is dominated by it. Superorganisms of copies might be economically - and possibly politically - unbeatable due to high productivity and coordination [Shulman 2010]. Over longer timespans this kind of monoculture might be significantly more vulnerable to evolutionary risks discussed below than a diverse population.

This scenario is of concern proportional to resource advantages: if more computing power equates to more practical power (hacking ability, economic power, political power, intelligence) and the cost of acquiring this is less than the gains, then runaway agents forming monopoly-like structures are possible. This could happen due to bad computer security and easy copying, or strong economies of scale for the emulations’ collective income.

A contest between emulations

If a large and diverse set of emulations exist, it may be rational for all of them to fear that one of the others will engage in sudden, unbounded self-reproducing behaviour. Especially if such a runaway emulation were "violent" — ie, willing to seize control of the computational resources of other emulations using malware, subterfuge, or physical force — every emulation might calculate that the entirety of its (extremely long) future life was at risk from such a possibility. Comparatively altruistic and non-sociopathic emulations might even decide to engage in drastic self-reproducing behaviour to head off such risks. Other emulations might interpret this

defensive behaviour as a threat.

Unless and until a stable political economy of computational resources is established, the risk of an escalating conflict between emulations exists. The risk exists if there are any runaway, rapidly reproducing emulations, or if comparatively well-intentioned emulations are unable to persuade each other of their good intentions.

The level of risk of this scenario may depend on some addressable questions about computer security. One of the most pointed threats that one emulation might pose to a second is that of using malware to seize control of the second emulation's computer. It is therefore useful to ask whether, in a hard-fought contest between one party defending a computer running an emulation, and another party trying to seize control of it, the structural advantage lies with the attacker or the defender.

Amongst computer security experts, it is commonly believed that attackers tend to win, because every system has many bugs, of varying levels of obscurity; the defender needs to fix all of them, while the attacker need only find one [Schneier, How Changing Technology Affects Security, *IEEE Security & Privacy*, March/April 2012]. This observation is an initial reason for pessimism about the ability of emulations to achieve security by defensive means.

The situation may not be so dire, however. One important method available to defenders is "defense in depth", a strategy in which defenders expect their systems to be vulnerable, and respond by trying to ensure that the attacker would have to compromise quite a number of subsystems at once in order to succeed completely. Ideally, some of these subsystems are hidden or wrapped within each other. Defense-in-depth may allow defenders to play offense for a while, with their objective being to detect an intrusion and end it, or trace it back to its source.

More concretely, emulations may be able to employ defense-in-depth techniques such as running copies of themselves that are not attached to the Internet, whose primary purpose is to monitor the hardware and software of their online copies, and reset and reinstall it if any unusual states can be detected.

We believe this kind of powerful defensive strategy might be enough to tip the scales to allow emulations to be run in defensively secure hardware. If this is not true, it can be argued that conflict between emulations is a serious risk if there are more than a small number of them, and that the consequences are fairly drastic [Shulman 2010]

Conflict between humans and emulations

It is possible that some humans beings will feel very threatened by emulations, especially if those emulations possess superhuman abilities, drive wages below human subsistence levels [Hanson 2010], or are unable to politically convince humans that they are harmless. Even the mere existence of emulations would no doubt outrage the sensibilities of people with philosophies incompatible with emulation. If groups of humans tried hard enough to restrict, slow or destroy emulations, it is possible that the emulations would fight back, and the conflict might escalate.

It has been argued that there are prudential reasons not to develop technologies allowing the emergence of enhanced humans or posthumans because of the risk of conflict [Annas GJ, Andrews LB, Isasi RM. Protecting the endangered human: toward an international treaty prohibiting cloning and inheritable alterations. *Am J Law Med*. 2002; 28(2-3):151-78.]. While much of the concern appears based on overconfident assumptions about inter-species psychology and relations, the emulation case gives at least some more concrete reasons to be concerned.

As noted by Allen Buchanan, enhancements that enable new and distinct forms of complex cooperation (in his case cognitive enhancement, in ours brain emulation and the possibilities it entails) could result in individuals with the same moral status but unequal rights. [Buchanan, Allen. 2009. Moral Status and Human Enhancement. *Philosophy & Public Affairs*. 37(4): 346-381.] This would occur due to the complex cooperators overarching the 'dominant cooperative framework' of the society by virtue of their abilities, and the more complex society that resulted would involve new rights and responsibilities the simple cooperators could not participate in. This would not necessarily have to be a moral catastrophe (consider the reduced rights of children or mentally disabled people), but a transformation of society that moved the previous majority from the center to the periphery would no doubt cause strong reactions.

Neuromorphic AGI as a side effect

Serious concerns about the safety of superintelligent artificially intelligent systems have been raised [Yudkowsky 2008, Omohundro 2008, Muehlhauser & Salamon 2012]. Many of these center on the risks posed by intelligent agents that can improve their capability (and hence power) more rapidly than societal structures can handle, and that have badly designed motivations. Constructing motivational systems that achieve desirable behaviors in real-world situations is hard, doing it for potentially self-enhancing and very powerful systems even more so. Emulations are potentially a solution in that they are harder to radically upgrade since they are based on the messy human neural architecture, and that their motivations are human-like since they are directly based on scanned human brains.

However, developing the necessary technology for brain emulation might provide cognitive science with enough useful data and impetus to invent neuromorphic AGI systems borrowing algorithms or subsystems from brains. [Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. St. Martin's Griffin 2005. Ray Kurzweil. H*ow to Create a Mind: The Secret of Human Thought Revealed*, Viking Adult 2012] This is supported by the fact that past AI research has benefited from techniques derived from biological discoveries, such as neural networks, reinforcement learning, genetic algorithms, perceptual hierarchies etc. Such AGI would likely be in the "messy" category of AGI systems where motivations and internal states are hard to analyse, yet potentially "clean" enough that rapid self-improvement is possible. Neuromorphic AGI might hence be about as risky as *de novo* AGI, and a possible side effect of emulation projects.

Speaking against neuromorphic AGI being an inevitable consequence of brain emulation is the fact that the bio-inspired contributions to artificial intelligence often required decades to emerge after being discovered among biologists. Knowledge diffusion across discipline boundaries

might be slow enough to reduce neuromorphic AGI risk. However, there is no guarantee these boundaries will remain as impermeable as they have been in the past.

A related possibility might be that brain emulations rapidly lead to radically enhanced or modified minds. It is easier to edit a software mind than a biological one, fast minds can be rapidly tested, software enhancements can be connected to the emulation using virtual interfaces, and methods of software optimization can be applied to improve performance. This could produce superintelligent or essentially non-human minds that are as problematic as *de novo* AGI, and have runaway potential.

Evolutionary risks of emulations

The transition into posthumans represented by emulations might potentially lead to evolutionary pathways that threaten value. For example, if emulations lack phenomenal experience or through 'upgrades' lose it, mindless but intelligent systems might gradually crowd out humans and other systems able to experience valuable experiences. Another possibility might be that competitive pressures lead to emulations optimized only for work, not for any enjoyment. [Nick Bostrom, The future of human evolution, in *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Ria University Press: Palo Alto , California , 2004): pp. 339-371. http://www.nickbostrom.com/fut/evolution.html] If emulations can invest their earnings into resources necessary for running further copies, the most prolific emulations may evolve towards spending all their resources in copying (since the individuals holding back will soon be outnumbered by the fastest replicators; see [Robin Hanson, Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization. 1998]), using up all available resources. It only takes one individual with the motivation of unbounded copying (for whatever reason) to trigger this process.

This form of existential risk is more subtle than the direct disasters described above. It might involve rational free actions of many agents or abstract evolutionary forces moving the world towards a dystopian outcome, with nobody being responsible for the emerging disaster or in a position to prevent it. Bostrom suggests the need of global coordination to handle this form of risk, but achieving the coordination is hard and sufficiently strong coordination systems might themselves produce dystopian outcomes. The specific risk can also be impossible to predict beforehand, like many emergent properties of complex systems.

Predictable Factors that Increase or Decrease Risk

In general, the possible circumstances under which human emulations might appear, and the manner in which the risks discussed above might play out, are far too complicated to characterise.

However, we believe that there are some very high-level categories for these circumstances that are well-defined and possibly even controllable in advance, that are deserving of analysis.

The core insight is that there are several technological ingredients to emulation, and these might fall into place in different orders.

How might emulation start?

There are at least three indispensable ingredients that would be required for human brain emulation to be conceivable:

1. a scanned copy of some individual's neural (and neurochemical) architecture;
2. a sufficiently good model of the computational mechanisms in the brain to be able to run the scanned copy; and
3. sufficient computational resources to run the emulation.

The historical circumstances under which an emulation might first occur can reasonably be grouped into 8 categories, depending on whether each of these prerequisites had existed for a long time or have recently been developed.

The simplest cases to analyse are those in which two of the three preconditions had been satisfied for a long time, and one was clearly the last to be met.

Scanning comes last

In this scenario, scanning of a human brain occurs last, but the models necessary to emulate neural systems are already well-developed, and available computational resources are already well beyond those necessary for emulation. Scanning might either be a hard problem or require an extensive technological infrastructure.

Under such circumstances, after the first person (let us call her Eve) to be scanned had been successfully debugged, many copies of her could be run, or a smaller number might be run at speeds much greater than those of physical humans.

The consequences of such a scenario would depend greatly on Eve's personality and talents; on the nature of the institution (or institutions) which have access to her scanned dataset; and on the length of the period of time before other individuals were scanned.

Depending on the institutions, Eve might or might not possess autonomy. She might also exist in several or many places, with different degrees of autonomy.

If Eve possesses autonomy and the wrong personality, there would be some risk that she might become a "runaway", exerting overwhelming and counterproductive influence on future events. This would require her to have some way of acquiring, through purchase or otherwise, a very large amount of computational resources before other humans were scanned and began to compete with her.

Even if Eve is herself unproblematic as an individual, if she becomes very prevalent she might become a “monoculture”: a large number of copies exist, gaining an incumbent economic and social advantage over later emulations, yet possibly having mental weaknesses that could come into play on a vast scale if discovered too late. Even the mere lack of diversity might have an adverse effect on society and promote some of the evolutionary existential risks mentioned above.

If Eve lacks autonomy she will be a potentially powerful tool for a group or agent; by exploiting the advantages of emulations they could use her to gain significant power. Other groups may be motivated to stop them using destructive means.

A model comes last

In this scenario, techniques for scanning the neural structure of human brains would be well developed, and computational resources would be plentiful, before anyone fully developed and debugged code to successfully run an emulated human.

At that point, it would become possible to run however many humans had been correctly scanned. Presuming that the algorithms necessary for the emulation were not kept secret, a number of emulated humans would appear during a short period of time, and vary in their institutional settings and levels of autonomy.This scenario allows for a significant speed or copy number advantage of the emulations, since plentiful computing power is available.

Unlike in the scanning last scenario, here the possibility of emulation might come as a surprise to society: there are no precursors, no precautions. Given past failures of emulation, many decisionmakers and the public might have concluded that the approach was a dead end. There would hence be little planning for this eventuality, including weak protections of autonomy or personhood.

This is similar to the analysis in [Shulman, Carl, and Anders Sandberg. 2010. "Implications of a Software-Limited Singularity." In ECAP10: VIII European Conference on Computing and Philosophy, edited by Klaus Mainzer. Munich: Dr. Hut. http://intelligence.org/files/SoftwareLimited.pdf] of how the role of software and hardware influences a future AI breakthrough: software-limited breakthroughs are likely to have hardware overhangs making them rapid and giving the software significant capabilities.

Computational resources come last

The number of operations per second required to emulate a brain in real time is unknown. Sandberg and Bostrom (tables 8 and 9) calculate a range of possibilities, parameterised by the level of physical detail that must be included in the calculations. These possibilities range from real time brain emulation being possible on existing supercomputers, to situations where Moore's law would have to continue for many decades to make real time emulation possible.

In this scenario demonstrations of small animal brain emulations are achieved first, but due to excessive computing requirements human scale emulations are infeasible, slow or tax available storage. The possibility of better emulations as hardware improves or emulation techniques

are improved is fairly clear to observers of the research. Society has some time to adapt to the prospect of emulations.

In this scenario arms races might occur between competing groups wishing to develop human-level emulations for their long-range strategic importance, but they would be limited by the overall state of hardware; the best they might do is to aim for faster special purpose hardware or improving the algorithmic performance of emulation software.

The earliest human scale emulations will not have a speed advantage over biological humans and will likely be located in major computing installations. This means that they do not hold any strong initial advantage over humans (economically or intellectually), that they will be few, that copying emulations likely requires copying and moving storage media, and that they are less likely to be subject to hacking that aims at making surreptitious copies or replacing one emulation with another one. However, given the mature state of neuroscience in this scenario, it might be possible that subtle forms of emulation hacking has been invented using animal emulations or neurocognitive insights.

Computer Security Considerations

Several of the risks discussed above depend a great deal on computer security considerations. While contests between people who want to break into computers and people who want to stop them are already of considerable economic significance [cites], the stakes in such contests might be much higher for emulations. If computer security cannot be guaranteed emulation autonomy cannot be guaranteed: copies can be stolen, altered and run by slavers. An emulation that controls more computers could think faster or make more copies of itself; an emulation that lost control of all of its computers would cease to exist at all. For emulations, warfare by hacking looks a lot like conventional warfare between states, but on vastly shorter timescales. It may therefore matter a great deal what the balance of power looks like in such conflicts.

We propose two competing hypotheses about the balance of power in computer security conflicts between emulations. Which of these is correct will turn out

Hypothesis: The Attacker Always Wins

A common belief among theorists and practitioners of computer security is that the task of defenders is vastly more difficult that the task of attackers. As Schneier put it,

> Security designers occupy what Prussian general Carl von Clausewitz calls "the position of the interior." A good security product must defend against every possible attack, even attacks that haven't been invented yet. Attackers, on the other hand, only need to find one security flaw in order to defeat the system. And they can cheat. They can collude, conspire, and wait for technology to give them additional tools. They can attack the system in ways the system designer never thought of. [Bruce Schneier, Security pitfalls in cryptography, *Information Management & Computer Security*, 1998 http://www.schneier.com/essay-028.html]

This dynamic appears to have played out in practice. We know of no examples of complex, security critical software systems that have remained invulnerable to attacks, while there are abundant examples of vulnerabilities being discovered in carefully designed defensive systems (consider the past cases of ssh, openssl, DRM, and crypto co-processor vulnerabilities).

If one vulnerability was all that was required for one emulation (or perhaps many emulations) to lose control of their computers, the implications would be troubling. A balance of power that favours aggressive behaviour could be expected to lead, one way or another, to aggressive behaviour. Many of the risks we considered previously, including conflicts between emulations, and conflicts between humans and emulations, would be greatly exacerbated if the Attacker Always Wins hypothesis were true.

Emulations would only be vulnerable to hacking to the extent that they were connected to a network; it might be possible for them to isolate themselves substantially from vulnerability by refusing to access networks without several levels of indirection. But such emulations might find themselves at a tremendous disadvantage compared to their cousins that were willing to take the risk of wiring a network connection deep into their emulation code.

Counterhypothesis: Emulation Systems are Defensible

There are some arguments that contradict the attacker-always-wins, winner-takes-all hypothesis. If these arguments persuasively refute the hypothesis, they may justify more optimistic conclusions about the political and economic stability of societies that include emulations.

The central observation in one such argument is that the physical control of hardware can constitute a decisive advantage for computer security defenders under some circumstances [Ref?]. The benefits of controlling the hardware are not unlimited — essentially, they comprise the defender's ability to turn the machine off, to perform certain kinds of forensics, and to install new software in place of the previous software. These powers are bounded by the defender's limited ability to know when she has lost control of the software on her machine. The crucial point is that, if the defender has a way to test that her machine is performing a task as she intended, she can use physical control decisively. If she cannot efficiently test that the machine is performing the task she wants, physical control is of much less use.

Some tasks are of this latter sort. For instance, it is extremely difficult to ensure that one's computer is not secretly retaining a copy of one's password somewhere, to be sent in encrypted form to an attacker at a later time. As a result, malware that steals passwords is extremely common and hard to defend against [cite]. By comparison, malware that steals a significant portion of the CPU cycles from a scientific supercomputer would be much harder to hide, because the owners of the machine have a reliable test (is the supercomputer still performing our astrophysics simulations?) and the resources to apply it. If emulation data is large compared to normal data traffic, then copying is detectable.

As far as belligerent emulations go, most of the resources they would be interested in procuring look more like the supercomputer case and less like the password-stealing case. Provided they

employ sufficient defense-in-depth, emulations could generally be safe from being hacked out of existence if they, their security systems, or their human allies were capable of (1) testing that the emulation was still running and (2) restoring it from a backup if it had been replaced, drastically altered or subverted.

One limitation to this counter hypothesis is that subtle alterations to the victim emulation might be much harder to detect. The opacity of brain function assumed in the WBE scenario works against detection, but may only partially help against developing “neural exploits”. If the attacker is capable of altering the victim's thought processes is specific ways, security systems might have tremendous difficulty in determining that such an attack had succeeded. Emulations change over time in complex ways due to experience (a process that cannot be removed without losing most of the value of brain emulation in the first place), so merely looking for any change will not work. Modified minds might communicate with their controllers using covert channels appearing as normal behavior, and software subroutines could be inserted in the low-level biological emulation. From a computer security perspective it is important to recognize that a brain emulation is a virtual machine that can execute near-arbitrary code.

Even if the neural integrity is not a problem it is worth remembering that many of the most successful security exploits have been aimed at the human component: either by “social engineering” manipulation of naive users, or persuading them to go along with the scheme. Emulations are by their nature potential targets, and it is unlikely they could be shielded from potentially malicious communication. If the emulations have control over their security they could be convinced to give it up, and if they do not they might still supply insider information and help. The problem is not too dissimilar from the normal insider security problem, and presumably about as hard to solve.

The openness of emulation technologies

The software and models for running emulations might exist in different states. At one extreme it might be completely open, with free/open source software codebases, and scholarly publications explaining the methods in detail. At another extreme, the technology might be completely hidden, existing only within certain organisations and carefully guarded against disclosure to outside parties. In between states are also possible: the code might be available in proprietary, closed-source form, or academic publications might disclose some but not all details of the emulation methods.

Scanned datasets of human brains, and the technologies for creating them, could also be openly available or privately held.

Datasets are likely more easy to keep private than emulation technology in general. There exists a large field of computational neuroscience, and solutions can be reverse-engineered or reinvented. However, brain datasets *cannot* be reverse-engineered (otherwise the AI problem would be solved and the situation would be completely different). To produce new datasets access to scanning is required (gaining access to suitable brains might also be nontrivial). Scanning technology might of course also be reverse-engineered, but this requires nontrivial engineering and construction in the physical world, something that is likely to be comparatively

slow. In scenarios where scanning is the bottleneck the existing hardware overhang will make the already existing emulations fast and numerous, making the reverse-engineering delay a major problem for competitors seeking to oust an incumbent group.

One significant consequence of open emulation technology would appear to be increasing the population of emulations that might exist at any one time. Open input datasets might cause the creation of numerous copies and mutations of a given emulation, with varying degrees of autonomy. Whether this more widespread creation and multiplicity of emulations increases or decreases risk seems to depend on the relative likelihood of different risk dynamics. For instance, a single emulation or emulation-controlling entity running amok and seizing catastrophic amounts of power seems much less likely if numerous brain datasets, and the methods to run them, were widely available. Wide access to emulation code and datasets might however not reduce geopolitical conflicts if they are driven by inequalities in ability to exploit the economical benefits of emulation. Conversely, proprietary technology by its nature involves an inequality in access, but some risk mitigation methods (see below) might depend on it.

The question of security of open software is complex and has been widely discussed [Marit Hansen, Kristian Köhntopp, Andreas Pfitzmann, The Open Source approach — opportunities and limitations with respect to security and privacy, Computers & Security, Volume 21, Issue 5, 1 October 2002, Pages 461-471]. Proponents argue that open source software defects are found and corrected quickly [Raymond, E. S. 1999. *The cathedral and the bazaar*. Available at http://www.tuxedo.org/~esr/writings/cathedral-bazaar/], although empirical data suggest that software quality might be comparable [Audris Mockus, Roy T. Fielding, James D. Hebsleb. Two case studies of open source software development: Apache and Mozilla. ACM Transactions on Software Engineering and Methodology. Volume 11 Issue 3, July 2002 Pages 309 - 346. O.H. Alhazmi, Y.K. Malaiya, I. Ray, Measuring, analyzing and predicting security vulnerabilities in software systems, Computers & Security, Volume 26, Issue 3, May 2007, Pages 219-228] In simple models, we may expect the practical security of software systems to be unaffected by whether the source code to those systems is open or secret; see [Ross Anderson, Security in Open versus Closed Systems -- the Dance of Boltzmann, Coase, and Moore, in *Proceedings of Open Source Software Economics*, Toulouse 2002. http://www.cl.cam.ac.uk/~rja14/Papers/toulouse.pdf] However, open source might be important in establishing the trustworthiness of software and hardware [Hansen et al. 2002]. This is of particular concern for emulations worried about their autonomy, and may help improve trust in the larger emulation ecosystem.

Conclusion

It is unclear whether the securability of software systems is a lever we can control or simply a fact of life in our mathematical universe.

However, practices of software engineering, maintenance and governance can influence how close to the maximum practically attainable security we get. Even if the defensibility hypothesis is true, very significant ongoing expenditures on software auditing and quality standards might be required to avoid the attacker-always-wins scenario.

Can Emulation Risks be Mitigated?

Given the arguments in the previous section that the emulation of humans would lead to certain kinds of serious risks, it is natural to ask whether anything can be done to mitigate these risks.

Open societies and institutions

Generically, democracies rarely wage war on each other. Open societies are resilient since they allow disagreements to be aired and resolved in a variety of non-violent ways. Insofar such democratic means of protection of rights and participation are extended to emulations, incentives for violent behavior are reduced. This requires protection against autonomy-reducing practices, since they both preclude real democratic participation and undermines democracy itself.

However, local systems that favor the safety and prosperity of emulations may also produce nontrivial geopolitical disagreements. If some countries treat emulations as citizens and others as mere software, they might gain very different economic growth rates and political outlooks due to the migration and copying of different emulations. Democracies have gone to war against regimes judged to be committing atrocities or preparing for war. An emulation-dominated democracy might be militant against countries treating their virtual brethren badly.

Does diversity increase or decrease safety?

An important structural question is whether having a larger number or a smaller number of emulations in existence early on would lead to safer or more dangerous outcomes. If there were a smaller number, we might be putting a lot of eggs in a small number of baskets, by hoping that we (and fate) chose the right personalities and the right institutional surroundings for the creation of a technology with potentially immense consequences. Having a greater diversity of emulations early on would tend to increase the risk that some of these emulations might wind up in conflict with each other, and the risk that evolutionary pressure might select for exceptional aggressiveness in some way. But, conversely, it would ensure that a wider range of cultures, personalities, and institution types could influence the transition.

One factor which might concretely affect this balance is whether computers can be made secure enough for emulations to live in them safely. If the answer was yes, the risk of disruptive conflicts amongst emulations or between emulations and humans might be lower. A “yes” answer might also decrease the risk of evolutionary pressure for aggressiveness, although not completely.

Avoid or delay emulation technologies

One way to mitigate emulation risks would be to prevent emulation from occurring at all.

There is little precedent to indicate that it is possible for humanity to get close to developing a particular new and important technology, but hold back from actually developing it. Two notable candidates are human cloning and land-based autonomous robotic weapons. Neither of these technological embargoes have long enough records to be regarded as durable.

It might be more possible to avoid the development of a technology like whole brain emulation by holding off on investment in precursor research such as the classification and characterisation of neuron types and the development of novel microscopy techniques. But such an endeavour would require a near-consensus throughout the developed world that human brain emulation is both possible and sufficiently dangerous to justify de-funding entire branches of science.

It is more likely that a serious effort to prevent human emulation would delay the development of the technology for some period of time. This leads to another question: would the serious risks posed by emulations increase or decrease if such a technology appeared at a later point in human history?

Several circumstantial variables might change over years and decades that make an event like the appearance of emulated humans more or less dangerous. One is that people might have thought more deeply about the possibility of emulation, and have formulated institutional proposals for reducing the likelihood of conflict. A second variable is the zeitgeist — at different times, people might regard emulations differently. Perhaps, for instance, during times of war emulations might be thought of as an excellent kind of cybernetic soldier. At other moments they might be thought of as perfect for achieving (post)human immortality, and so forth. A third variable is the state of development of other relevant technologies during the era when emulations might come to pass. For example, advanced nanotechnology may allow rapid customized manufacturing useful for extending the hardware base and having short product cycles very suited for an emulation economy.

It seems that the first two variables — evolving social understanding of emulations, and the nature of zeitgeist — are too complicated to be predicted or characterised in any useful way. The third, however, may be amenable to some level of prior analysis.

Even if technologies are hard to delay, it might be possible to influence the ordering of technologies. For example, funding of technologies that might safeguard against one or more risks might increase the chances of them arriving before brain emulation. Support for neuroscience research and computational modeling can make it more likely that the hardware limited scenario comes to pass than the scanning or neuroscience limited scenario. Unlike slowing down emulation technology, such research pushes might be motivated on their own grounds, not requiring widespread consensus.

A similar case can be made for work on improving computer security. There are already many good reasons to push for it, and improvements likely reduce some of the emulation risks down the road. Even if software insecurity is fundamentally unavoidable, having better understanding of the risks and how to live with them might allow better emulation policies. The acceleration does not have to be general: if key software types or functions for emulation (for example, high performance computing or cloud security [Waleed W. Smari, Luca Spalazzi, Yacine Zemali, Recent developments in high performance computing and security: An editorial, Future Generation Computer Systems, Volume 29, Issue 3, March 2013, Pages 782-787. Mark D. Ryan, Cloud computing security: The scientific challenge, and a survey of solutions, Journal of Systems and Software, 2013]) can be identified beforehand they can be targets of focused effort.

Other technologies that might impact the safety of emulation are surveillance related. Global surveillance systems might reduce the chance of undiscovered secret large-scale projects. More local surveillance may reduce the chance of defection by individual agents, whether software or biological. Emulations themselves can be placed under the most through surveillance possible. Effective deception detection - potentially a spin-off technology from emulation work - might allow ways of establishing trust among agents. However, whether these possibilities increase or decrease risk sensitively depends on how and for what purpose they are used, whether there is appropriate oversight and accountability. In this case know-how (governance or technology) that reduce the risk from the technology itself needs to be accelerated compared to the surveillance power, even if that power correctly used would reduce emulation risk.

Mitigate inequality effects

Allen Buchanan's ethical analysis of rights in a society of enhanced and unenhanced people concludes that human rights are not rendered obsolete by the existence of posthumans (and moral status of unenhanced is not reduced). Even if the enhanced dominate the cooperative framework and enjoy some additional rights due to their abilities, this can be in the best interest of both groups. This has some similarity to Rawls' difference principle, where inequalities are permissible if they give the greatest benefit to the least advantaged members of society. [John Rawls, *A theory of justice*, Belknap 1971]. Hence if the radical differences induced by brain emulation can be made to improve the conditions of the least-advantaged they can be just.

Stuart Armstrong has suggested that groups working on brain emulation might set up funds for the benefit of the rest of humanity, for example that 1% of the possible profit will be shared by everybody else. This would help build trust and possibly avoid the most extreme inequality effects, but perhaps most importantly serve as a strong signal about the expected potential of emulations in a scan- or model-limited scenario. [Stuart Armstrong, World funds: implement free mitigations, November 10 2011. http://blog.practicalethics.ox.ac.uk/2011/11/world-funds-implement-free-mitigations/]

Similarly precommitment to set up a dominant cooperative framework to guarantee rights in the case of brain emulation (for example redefining personhood to be substrate independent) might also help fix ethical/legal inequality and not just economic inequality.

A more general mechanism would be to reduce the dramatic winner-takes-all aspects. Ensuring that systems able to run emulations are secure enough that sudden takeovers are unlikely would reduce the potential for conflict. How security scales with the size of installations (c.f. [Ryan 2013]) might create natural incentives or disincentives for centralized control over the emulation infrastructure.

Making it costly for single groups to own or rent massive computing might also reduce the power held by them in monopoly or oligopoly scenarios. This could be implemented through progressive taxation of CPU cycles, lower bounds on emulation social spending, or other forms of taxation or administratively imposed diseconomies of scale. However, given the fluidity of emulation labor there is a certain risk that this would just prevent the most ambitious players from setting up their data centers in countries with a strong regulatory climate: global agreements are likely necessary to reduce this risk.

Global agreements

Agreements on proper treatment of emulation security, ethics of emulation research and application, and the rights and responsibilities of emulations can influence risks.

For example, regulation of copying or rules that reduce the incentives for rapid copying can reduce the risk of dystopian economic scenarios. This might include bans on multiple instancing beyond backup copies, requiring splitting of funds between autonomous copies or minimum wage payments to non-autonomous copies. Anti-trust legislation might be applicable to large clades of emulations.

The weakness of this approach is the complexity of reaching widespread agreement, especially when emulation research or application might be hidden, and the benefit of defection (for groups or individuals) can be very high.

To make matters worse, it is likely that emulation technology is going to appear science fictional in scenarios where successful modeling occurs late, giving little time to reach agreement before the technology becomes operational. After it has been proven the need for agreement might be apparent, but the emulations are likely to be fast and hence likely to produce significant social strain before most agreements can be reached. In slower scenarios (especially ones limited by computing power) there is more time to reach an agreement before or after human emulations occur.

Agreements are only half of the work: the crucial issue is whether effective enforcement can be implemented. A legal framework without effective enforcement might be just window dressing.

Typically enforcement requires detectability (the activities to be handled are detectable by the authorities), enforceability (states or other signatories can execute the agreement), arbitration (parties can appeal or arbitrate), and transparency, all aimed at deterrence of unwanted activity.

Detectability of emulation-related activity depends on whether the hardware needed is easy to hide, and whether particular software activities are detectable. If scanning is a large-scale special purpose activity scanning could be detected and controlled, at least regulating the inflow of brainscans into the emulation infrastructure. Similarly large scale computing might be a natural bottleneck that allows inspection of emulation practices. But if these become irrelevant because scans are copied or computers able to run emulations are ubiquitous the only remaining option is the ability to detect what occurs in software. Historically this has not been very successful.

Limitations of detectability are likely to dominate the problems of enforcability, but they may be complex on their own. Trusted emulation architectures may have backdoors for legal inspection and intervention, introducing a potential security weakness and thick legal issues about due process. But emulations running on non-trusted architectures might be far more relevant as victims or perpetrators, requiring significant cyber-policing agreements.

Conclusions

If brain emulation does one day come to pass, it will be up to the people (biological and emulated) of that era to find a way to make the transition from a planet with one intelligent species to a planet with at least two quite different intelligent species, as peacefully and as wisely as possible. They will have a far better perspective on the problem than we can achieve today, and we can only imagine a small fraction of the methods they might attempt to use.

For the time being, it merely seems worth asking whether there are policies and institutional arrangements we can pursue that would set up the pieces on the board so that the game, if it is played, is easier. Should we be prioritising funding to accelerate brain scanning projects, or withholding it to slow them down, letting Moore's law continue in the meanwhile? Should we be building institutions to encourage the open-sourcing and open-sciencing of relevant projects, or should we prioritise funding them carefully, behind veils of confidentiality? Should we want these projects conducted by militaries, corporations, or academia?

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	12	11	15	13
CPU,scan,model	13	14	20	12
scan,CPU,model	15	10	17	11
model,CPU,scan	12	10	14	11
model,scan,CPU	4	3	4	3
scan,model,CPU	3	2	6	2

This paper can only give tentative answers, but a few patterns stand out in the scenario combinations. The technology ordering likely dominates over questions of openness and security in terms of risk contribution: the slower scenarios are safer and less likely to lead to extreme, low-autonomy situations. Insecurity coupled with rapid breakthroughs has potential for tempting agents into risky behavior. The openness of code might be more relevant in terms of trust of the emulation infrastructure, but this trust is also important for establishing trust in the institutions of the emulation world.

This paper makes one core conjecture: that it would be best to press ahead quickly with the neuroscience and microscopy projects necessary for emulation, in order to reduce the probability that emulations appear suddenly and dramatically.

The paper also identified several open questions that appear to be important in predicting the possible political economies of societies that include brain emulations. One was the question of whether emulations' physical control of their own hardware will be sufficient to generally provide them with protection against network-based attacks, or not. Another was the question of whether emulations will be able to tell if they have lost their own autonomy. For each of these questions, answering "no" appears to make the conjecture above more likely to be true and more important.

Acknowledgements

Thanks to Toby Ord, Stuart Armstrong and Nick Bostrom for many helpful discussions.

ATTIC

(Material moved out of the article for now).

Anders' copy

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	Single group control likely. Potential for non-autonomous slavery giving group significant power. Pre-emptive attacks against it possible.	Single group control likely. Potential for non-autonomous slavery giving group significant power.	Multiple versions of Eve with varying autonomy.	Multiple versions of Eve with varying autonomy.
CPU,scan,model	Surprise to society, hardware overhang, potentially several original scans. Pre-emptive attacks against leading group(s) possible. Much depends on whether model knowledge is disseminated. Big strategic advantages for pre-emptive seizing computational resources.	Surprise to society, hardware overhang, potentially several original scans. Leading group has strong advantage by likely having monopoly of modelling method. Can turn this into more permanent advantage by buying resources to extend their base.	Surprise to society, hardware overhang. Many copies likely with varying degrees of autonomy. Big strategic advantages for pre-emptive seizing computational resources.	Surprise to society, hardware overhang. Many copies likely with varying degrees of autonomy depending on scan availability.
scan,CPU,model	Surprise to society, medium overhang, numerous scans. Depending on model dissemination monopoly or oligopoly on em technology.Model secret likely target	Surprise to society, medium overhang, numerous scans. Depending on model dissemination monopoly or oligopoly on em technology.	Surprise, medium overhang, numerous scans, varying groups using them. Risks of hacking between smaller groups.	Surprise, medium overhang, numerous scans, varying groups using them.

	for hacking in case of monopoly.			
model,CPU,scan	Like CPU, model, scan but less extreme	Like CPU, model, scan but less extreme.	Like CPU, model, scan but less extreme	Like CPU, model, scan but less extreme
model,scan,CPU	Limited overhang, few copies, time to adapt (chain from animals to humans). Hard to steal full copies due to limited computing.	Limited overhang, few copies, time to adapt (chain from animals to humans).	Limited overhang, few copies, time to adapt (chain from animals to humans). Much dependent on initial copies and reactors to them. Multipolar.	Safe. Multipolar scenario.
scan,model,CPU	Limited overhang, multiple copies, time to adapt. Hard to steal full copies due to limited computing.	Limited overhang, multiple copies, time to adapt. Likely centralized to major computing centers or groups.	Limited overhang, multiple copies, time to adapt. Various speed-limited versions, hacking or attacks possible.	Limited overhang, multiple copies, time to adapt. Various speed-limited versions.

Peter's copy

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	most danger (single runaway emulation, depending on personality)	most danger (single runaway emulation, depending on personality)	most danger (single runaway emulation, depending on personality)	most danger (single runaway emulation, depending on personality)
CPU,scan,model	some danger (conflict among emulations unless carefully managed)	drastically transformative, strongly organisation-dependent	most danger (likely conflict amongst emulations)	drastically transformative but diverse
scan,CPU,model	some danger (conflict amongst emulations unless carefully	highly transformative, strongly organisation-	most danger (likely conflict among	highly transformative but diverse

	managed)	dependent	emulations)	
model,CPU,scan	some danger (single runaway emulation)	some danger (single runaway emulation)	some danger (single runaway emulation)	some danger (single runaway emulation)
model,scan,CPU	low danger (structurally dangerous but gradual)	low danger	low danger (structurally dangerous but gradual)	least danger (gradual, diversity of emulations)
scan,model,CPU	low danger (structurally dangerous but gradual)	less danger	low danger (structurally dangerous but gradual)	least danger (gradual, wide diversity of emulations)

Specific sub-risks (When we disagree let's list Anders/Peter values), every value is 0,1,2,3.

Risk of em-em conflict

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	0/1 1	1/0 1	2/2 4	1/0 1
CPU,scan,model	0/2 2	1/0 1	3/3 6	1/0 1
scan,CPU,model	1/2 3	0/0 0	1/3 4	0/0 0
model,CPU,scan	0/1 1	0/0 0	0/2 2	0/0 0
model,scan,CPU	0/1 1	0/0 0	0/1 1	0/0 0
scan,model,CPU	0/1 1	0/0 0	0/2 2	0/0 0

Risk of (SUDDEN) evolutionary selection for aggressive emulations or em-containing organisations

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	1/1	1/0	1/2	1/1
CPU,scan,model	1/2	1/1	2/3	1/2
scan,CPU,model	1/2	2/1	1/3	1/2
model,CPU,scan	0/1	0/0	0/2-	0/1
model,scan,CPU	0/0	0/0	0/0	0/0
scan,model,CPU	0/0	0/0	1/1	1/0

Risk of surprise / human-em conflict

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	2/2	2/1	2/2	2/1
CPU,scan,model	3/3	3/2	3/3	3/2
scan,CPU,model	3/3	3/2	3/3	3/2
model,CPU,scan	2/1	2/0	2/1	2/0
model,scan,CPU	0/0	0/0	0/0	0/0
scan,model,CPU	0/1	0/0	0/1	0/0

Risk of monopolies (economic, future), by a single em-containing organisation, including single runaway

	closed,insecure	closed,secure	open, insecure	open,secure

CPU,model,scan	3/3	3/3	2/2	2/2
CPU,scan,model	2/2	3/3	2/1	2/1
scan,CPU,model	2/1	3/2	1/1	2/1
model,CPU,scan	3/2	3/3	2/1	2/2
model,scan,CPU	1/1	1/1	0/0	0/0
scan,model,CPU	1/0	1/1	0/0	0/0

Risk of unsafe neuromorphic AGI

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	0/1	0/1	1/2	1/2
CPU,scan,model	0/0	0/0	0/0	0/0
scan,CPU,model	0/0	0/0	0/0	0/0
model,CPU,scan	1/1	1/1	2/2	2/2
model,scan,CPU	1/0	1/0	2/1	2/1
scan,model,CPU	0/0	0/0	1/0	1/0

GRAND TOTALS

	closed,insecure	closed,secure	open, insecure	open,secure
CPU,model,scan	12	11	15	13

CPU,scan,model	13	14	20	12
scan,CPU,model	15	10	17	11
model,CPU,scan	12	10	14	11
model,scan,CPU	4	3	4	3
scan,model,CPU	3	2	6	2

Emulations can achieve superintelligence by mimicking human organisational forms

One prediction about AGI which is sometimes debated is the idea that AGI will be "more intelligent" than human beings. We argue that, at least for some important pragmatic definitions of intelligence, the answer is certainly "yes".

Definition 1 : A pragmatic definition of intelligence

One central aspect of intelligence is the ability to solve problems posed in an informational form. We say that entity A is *unambiguously more intelligent* than B iff for any information problem x solvable by A, B is also able to solve problem x, and there exists at least one problem y solvable by B and not solvable by A.

We say that entity A is *in some sense more intelligent* than B if there exists a problem z solvable by A and not solvable by B.

Proposition 1 : human organisations can be unambiguously more intelligent than individual humans

Proof: consider a human being Alice. We will show an organisation to be unambiguously more intelligent than Alice.

The organisation consists of a person named Jane, who is like Alice in all respects, except that when she realises that she has expended great effort and not succeeded in solving a problem, she can pick up her phone and call on her extensive team of well-paid and highly motivated colleagues for assistance. Jane's colleagues include scientists, engineers, linguists, pyschologists, historians, philosophers and con-artists.

If Alice is able to solve a problem x, Jane can solve it too. But there exist some problems y which are outside of Alice's domain of expertise, or which require more human labour to solve than Alice is capable of. For some problems y, Jane's colleagues are able to solve y. It follows that Jane and her colleagues are collectively more intelligent than Alice.

For instance, human organisations have demonstrated their ability to design from scratch jumbo jets which fly safely. It is doubtful that any individual human could design a jumbo jet from scratch.

Lemma 1 : emulated humans have the same non-physical capabilities as human organisations

Because the entire state of emulated humans is digital information, these entities may be paused, saved, copied within one computer system or from one system to another (subject to sufficient memory and network resources). Saved or copied emulations may be reinstantiated and executed in parallel. In addition to these basic but remarkable properties, systems emulating humans can perform a number of other "organisational" operations:

- Creating two or more copies of an emulated person, and having them work on the same problem in parallel. In many cases, overall progress on the problem will be at least the maximum of the agents' individual progress.
- Creating two or more copies of an emulated person, and having them work on the problem together.
- If more than one scanned human is available, the operations above can be performed with copies of Alice, Jane and any other agents present.
- Most humans experience moods and mental states during which their productivity rises and falls. For many problems, it is helpful to be in a focused mood. Emulations can be paused and saved in such moods, and for finely grained problems, the emulations can be instantiated to work on each task while in their most productive states.