The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes

Daniel Eth
Department of Applied Physics, Yale University, New Haven, CT, USA
E-mail: Daniel.eth@yale.edu

Keywords: AGI, *de novo* AGI, neuromorphic AGI, Whole Brain Emulation, nanotechnology, AI safety

Received: August 31, 2017

In this paper, we contrast three major pathways to human level AI, also known as artificial general intelligence (AGI), and we investigate how safety considerations compare between the three. The first pathway is de novo AGI (dnAGI), AGI built from the ground up. The second is Neuromorphic AGI (NAGI), AGI based loosely on the principles of the human brain. And third is Whole Brain Emulation (WBE), AGI built by emulating a particular human brain, in silico. Bostrom has previously argued that NAGI is the least safe form of the three. NAGI would be messier than dnAGI and therefore harder to align to arbitrary values. Additionally, NAGI would not intrinsically possess safeguards found in the human brain – such as compassion – while WBE would. In this paper, we argue that getting WBE first would be preferable to getting dnAGI first. While the introduction of WBE would likely be followed by a later transition to the less-constrained and therefore more-powerful dnAGI, the creation of dnAGI would likely be less dangerous if accomplished by WBEs than if done simply by biological humans, for a variety of reasons. One major reason is that the higher intelligence and quicker speed of thinking in the WBEs compared to biological humans could increase the chances of traversing the path through dnAGI safely. We additionally investigate the major technological trends leading to these three types of AGI, and we find these trends to be: traditional AI research, computational hardware, nanotechnology research, nanoscale neural probes, and neuroscience. In particular, we find that WBE is unlikely to be achieved without nanoscale neural probes, since much of the information processing in the brain occurs on the subcellular level (i.e., the nanoscale). For this reason, we argue that nanoscale neural probes could improve safety by favoring WBE over NAGI.

Povzetek: Analizirane so tri poti za dosego splošne inteligenca tipa človeške inteligence, poleg tega so analizirane potencialne nevarnosti in problemi.

1 Introduction

Scientists disagree about when humanity will develop artificial intelligence that is at least as smart as humans in most or all facets of intelligence, with common estimates ranging throughout the 21st century [1]. There's little disagreement, however, that such so-called *artificial general intelligence* (AGI) will be transformative.

The human species has used its high intelligence to influence the world more than has any other species, and an even greater intelligence in AGI could potentially influence the world even further. Many scientists therefore expect the creation of AGI to be the single most impactful advent in human history [2].

Consequently, there has been an increase in research into how to align AGI with human values (so that this impact is for the better) [3]. Most of this research focuses on a hypothetical AGI that's programmed from the ground up (de novo), and this de novo AGI (dnAGI) is often considered as an extension of existing machine learning research or as some other abstract utility-maximizing agent (such as AIXI) [4][5].

While creating dnAGI is one potential path to AGI, there are other paths as well. Comparatively little research has been performed investigating the risks and benefits from various avenues, despite the fact that each avenue poses different challenges.

In this paper, we investigate the major technological landscape leading to AGI, and we assess which technological trends appear likelier to favor positive and negative outcomes.

2 Three major paths to AGI

The three major paths to AGI are dnAGI, Neuromorphic AGI (NAGI), and Whole Brain Emulation (WBE). For dnAGI, computer programmers conceive of algorithms that yield intelligence. For NAGI, the human brain is studied, and certain key features of the brain's architecture are appropriated, yielding an intelligence with some similarities to human brains. For WBE, the brain of a particular human is scanned, this scan is translated into a model, and the model is run on a computer – yielding an intelligence similar to that of the person whose brain was scanned (the human is said to have been "uploaded").

Other paths to AGI exist, but for this paper, we will consider just these three. In many cases, insights about other paths can be inferred from insights about these three. For example, one other path is simulating a generic human brain instead of a specific one, and this sits somewhere between NAGI and WBE.

In *Superintelligence*, Bostrom posits that NAGI is the most dangerous of the three [6]. The logic here is straightforward and sensible. NAGI is much more "messy" than dnAGI and thus would be harder to align sufficiently with human values. WBE, while perhaps even messier than NAGI, inherently contains safeguards that NAGI and dnAGI do not – such as compassion and other human values (to the extent that the human being uploaded holds "human values").

It is still an open question whether WBE or dnAGI is safest. Bostrom initially appears ambivalent about this topic, but later implies that he'd prefer dnAGI. His main argument is that dnAGI is ultimately the most powerful kind of AGI, so humankind must undergo a potentially dangerous transition with the development of dnAGI, even if WBE had been developed beforehand. If instead dnAGI is developed first, WBE may still be developed later. But because dnAGI ultimately will be more powerful, we'll only face a risk from the first transition [6].

This argument presupposes that an advanced form of dnAGI would be more powerful than an advanced form of WBE. The architecture of the human brain fundamentally places a constraint on the capabilities of WBE, and this constraint does not exist for dnAGI. Tweaks may allow either of these technologies to reach a higher level of intelligence than that of any biological human, but the upper limit is presumably higher for dnAGI than for WBE, and the path of improvements is likely steeper for the less constrained dnAGI.

While we think there is some merit to this argument, ultimately getting WBE first may still be preferable. We would be remiss not to consider, however, that pursuing WBE might lead to the particularly unfortunate outcome of NAGI being developed first. Having said that, it is argued in this piece that it would be good to accelerate the development of WBE, insofar as this can be done in a manner which accelerates the development of WBE relative to that of NAGI to a significant degree (such that the chances of achieving WBE first increase, the chances of achieving dnAGI first decrease, and the chances of achieving NAGI first either decrease or stay the same).

Consider just how hard it may be to align a dnAGI to human values. Not only would we have to figure out how to align dnAGI to arbitrary values, but we'd also have to figure out how to specify human values in a manner the dnAGI would understand. How do you explicitly specify values such as fairness or happiness? This is prima facie a Herculean task.

Compared to the task of aligning dnAGI to human values, safely traversing the path to WBE seems relatively easy. If built correctly, WBE would be generally safe – even in the absence of significant work on AI safety. For dnAGI, on the other hand, this is not necessarily the case, and even large efforts specifically focused on AI safety might fail. Additionally, there is reason to believe that mistakes in WBE wouldn't be as dangerous as mistakes in dnAGI. For WBE, minor mistakes may be tolerable, since the brain is resilient to small perturbations (there is no reason to expect such a safeguard in dnAGI). Also, screening WBE for safety would be much easier than screening dnAGI for safety, since we have the fields of psychology and psychiatry that may help us diagnose antisocial tendencies in WBE.

Even if a WBE was unsafe, that situation itself would be much less dire than an unsafe dnAGI. It seems unlikely that the initial arrival of WBE will mark an immediate artificial intelligence "takeover" – after all, we already have 7 billion beings with human level intelligence, and no one has been able to accomplish such a takeover. On the other hand, the first arrival of dnAGI might be vastly more capable than the smartest humans and might quickly take over. Either WBE or dnAGI could iteratively improve its own code, leading to an intelligence explosion. For dnAGI, this "intelligence takeoff" scenario (which must not be confused with the separate event of a "takeover," previously mentioned) could be very fast, while for WBE, it seems quite unlikely that the takeoff would be anywhere near as fast. The messiness of the brain's architecture and the constraints of the brain may limit how quickly the intelligence of a WBE could be improved. Since WBEs would have a harder time quickly taking over or cognitively taking off, humanity would likely be able to pull the brakes on a dangerous WBE in a way that we might not be able to with a rogue dnAGI.

Even if WBE is developed first, the subsequent shift from WBE to dnAGI would likely be a much bigger risk than the initial shift to WBE. Therefore, the major risk associated with the development of WBE is likely not from the shift to WBE itself, but in how WBE would affect the shift to dnAGI.

If we upload humans who are particularly intelligent and ethical, the resultant WBEs would be the very agents we would want to work on the problem of creating dnAGI safely. One reason to in general expect the uploading of humans that are at least relatively ethical is that more people would probably want to upload people significantly more ethical than themselves than would want to upload those significantly less ethical than themselves – especially given the stakes. By no means should we assume that only ethical people would ever be uploaded, as unethical people might have the means to get themselves uploaded. But – especially when WBE technology is new and there is likely to be much public debate about who should be uploaded and how they should be chosen – it is likely that on average WBEs would be more ethical than the general population as a whole.

With enough hardware, the WBEs would be able to think much faster than biological humans. Since modifications in WBEs would be much easier than modifications of biological humans, WBEs could self-modify more than could biological humans. Every time a modification was introduced, there would be a risk of value drift, but WBEs would presumably want to avoid

modifying themselves in ways to become the kind of agents that they didn't want to become. If the WBEs were taken from diverse cultures, a well-coordinated group of such WBEs would begin to embody what Yudkowsky has dubbed "coherent extrapolated volition" – an idea that AGI would be best to do what humanity wanted, if we "knew more, thought faster, were more the people we wished we were, had grown up farther together; … where our wishes cohere rather than interfere; [etc]."[7]

In *The Age of Em*, Hanson uses well-accepted, academic science (both hard science and social science) to predict the broad strokes of a potential future world dominated by WBEs [8]. His analysis is useful for our purposes of evaluating dnAGI created by WBEs, as we can consider how the world Hanson describes could affect dnAGI safety concerns. Hanson's analysis focuses on a period when WBE technology has advanced to the point that renting WBEs for labor is generally cheaper than paying biological humans for the same labor. Market forces cause the number of WBEs to increase rapidly, and WBEs perform almost all labor previously performed by biological humans [8]. Since the primary driving forces in this scenario for the creation of WBEs (either by uploading new humans or copying existing WBEs) are economic, WBEs are selected to be particularly profitable workers. Additionally, training for WBEs and mind "tweaks" would be selected for their effects on profitability [8]. In this scenario, previous considerations about WBEs being particularly ethical due to the desire to upload particularly ethical people may not hold; market forces would presumably present a much stronger selection effect.

Hanson's application of economics, sociology, and psychology (among other disciplines) leads to many conclusions about how WBEs might live [8]. Most of these conclusions aren't obviously related to the safety of subsequently developed dnAGI, but some are. Hanson argues that compared to most biological humans today, we should expect most WBEs to be smarter, more rational, more work-oriented, more mindful, more patient, and less mistake-prone [8]. All of these traits imply a lower chance of making a mistake in AI alignment. Additionally, Hanson argues that we have at least weak evidence to expect most WBEs (again, compared to biological humans today) to be better coordinated, more law-abiding, more trustworthy, and more expectant of and focused on preparing for big disasters [8]. These traits seem to imply WBEs would have a greater chance of successfully coordinating to prevent the creation of an unsafe dnAGI. On the other hand, the world Hanson describes is one that is much more economically competitive than our current world, which would perhaps increase the chance of a tighter race for the development of dnAGI [8]. This competition plausibly could lead firms to neglect important safeguards in dnAGI that they might consider to be luxuries they could not afford.

Implicit in Hanson's analysis is an assumption that WBE will be achieved long before dnAGI otherwise would have been achieved [8]. After WBE is developed, the cost for WBEs would need to fall below the cost of biological human labor for most jobs, and then economic equilibrium would need to more or less be established – all without dnAGI or NAGI being developed in the meantime (even with WBEs working towards creating these other forms of AGI).

If WBE is the first type of AGI created, it remains to be seen whether or not the economic conditions necessary for the Hansonian scenario will be realized before other forms of AGI are developed. Either way, shortly after the development of WBE, many WBEs will likely be significantly more competent than most biological humans – at least on many matters relevant to dnAGI safety.

Would a team of such WBEs be able to create a safe dnAGI? Possibly. But if such a highly competent team cannot, it's even less likely that biological humans could solve the problem.

Furthermore, such WBEs may better enable dnAGIs to be taught human values. Some proposals for specifying human values do not involve explicitly specifying the values, but instead involve allowing AI to learn the values from humans. For instance, AI could theoretically glean human values by observing humans, and then the AI could further be trained with feedback on its behavior provided by humans [9]. One limitation with this approach is that the amount of data and feedback that could be produced in any timeframe would be limited. More data and feedback could be produced with WBEs, since WBEs could be run to think much faster than biological humans (Hanson has estimated that WBEs might generally think about 1,000 times faster than biological humans, and perhaps sometimes even 1,000,000 times faster or more) [8].

Another idea has been to interconnect AI into our nervous systems. In this scheme, humans would more or less act as the limbic system for an AGI that would carry out our wishes [10]. This approach has several limitations, and again WBEs would be quite helpful. One major limitation is the difficulty in physically integrating AI with the processing of the brain. This would surely be much easier to do with a virtual brain than in a biological brain. Another difficulty is that such human-AI systems would be limited in speed by the human thinking, and other, separate AGI would surely be faster. Again, WBEs – operating much faster than biological humans – might be fast enough for this proposal to actually work.

While it's possible that NAGI might be developed after WBE (and before dnAGI) and that NAGI may also be more powerful than WBE, similar arguments to above apply for why developing WBE before NAGI would likely be safer than developing NAGI without first developing WBE. It must be acknowledged, however, that the existence of WBE might make NAGI easier due to their similarity. On the other hand, insofar as this would be a dangerous path to follow, smart WBEs might be able to avoid it. Additionally, even if NAGI could be achieved sooner than dnAGI, the fast speed of thinking in WBEs might decrease the cost (in time) of simply waiting for dnAGI without first developing NAGI.

Taking all of the above into consideration, plotting the expected safety of different types of AGI versus their similarity to a human brain yields a J-curve (Figure 1).

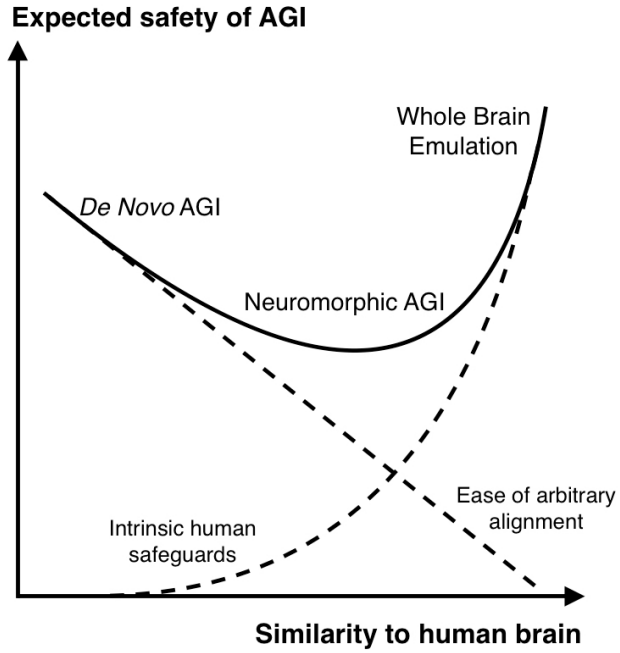

Figure 1: Expected safety of different types of AGI versus their similarity to the human brain. The downward sloping dotted line represents the fact that an AGI more similar to the human brain will tend to be messier and harder to align to arbitrary values. The upward sloping dotted curve represents the fact that an AGI especially similar to the human brain intrinsically contains human values. Taking these two effects into account reveals the solid curve to be a J-curve.

3 Technological landscape

Naively, we should work towards creating the type of AGI that would be safest if completed first. However, we must additionally consider interplay between different technologies. For example, while WBE may be safer than dnAGI, technology that progresses us towards WBE will often also progress us towards NAGI. This line of reasoning has led Bostrom to posit that *even if* we consider WBE to be safer than dnAGI, it *still* might be preferable to promote dnAGI so as to avoid NAGI [6].

This is reasonable as a general rule. When looking at specific technological trends, however, it is useful to consider the broader context of other relevant technological trends. In this section, we first will attempt to elucidate the major technological trends leading to each of the three major types of AGI. We will then combine these trends into a broader technological landscape and argue which trends are best to advance from an AI safety perspective.

3.1 *De Novo* AGI

The two major enabling technologies for dnAGI are computational hardware and what we will simply call "AI research," meaning certain types of software research (such as machine learning) that may be used in dnAGI (but not necessarily directly in WBE/NAGI). Much of current AI research doesn't progress us closer to dnAGI, and very little AI research is performed explicitly to direct us to AGI. In this paper, we're using the phrase "AI research" to mean any of this research that does progress us in the direction of dnAGI, whether or not the research is being performed for that purpose. Computational hardware and AI research, taken together and at an advanced enough level, seem necessary and sufficient for dnAGI. Interestingly, advances in both hardware and in AI research should lead to advances in the other one of the two. Improvements in hardware can lead to improvements in software for a variety of reasons, including allowing for more rapid testing of algorithms, and greater use of computationally heavy methods of algorithm design, such as genetic algorithms (which use Darwinian pressures to design and select algorithms according to certain criteria) [11]. Improved AI algorithms can be used to find superior computer chip designs. Hardware itself has a positive feedback loop, as greater computational capabilities are useful in powering the algorithms that help design chips.

Further upstream, nanotechnology research is a major enabler of improved hardware – especially as we reach the limits of silicon devices and other materials are needed to take over (options for such materials includes carbon nanotubes). Since one major subfield of nanotechnology is computational nanotechnology (using computers to advance nanotechnology, such as by simulating materials on the nanoscale), improved hardware and AI software would also aid nanotechnology research.

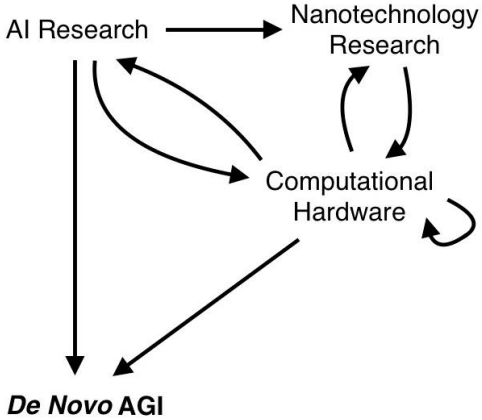

Figure 2: Technological pathway to *de novo* AGI. AI research and computational hardware are the main technological requirements.

3.2 Neuromorphic AGI

For NAGI, the major enabling technologies are hardware and neuroscience. With enough of the right kinds of neuroscientific knowledge to create algorithms of intelligence, and enough hardware to run such algorithms, NAGI could be achieved. Hardware was already discussed in the section on dnAGI. Since

computational neuroscience is a major aspect of neuroscience, both hardware improvements and AI research would also aid neuroscience (in much the way that they'd also aid nanotechnology). Nanotechnology could also provide much benefit for neuroscience through the creation of nanoscale neural probes. Such probes would have many benefits for neuroscience over existing brain scanning technologies – in particular, they could scan the brain with subcellular resolution, *in vivo*, and many could potentially be used in parallel to determine the architecture and function of neural circuits. Advances in neuroscience would additionally be useful for designing such probes.

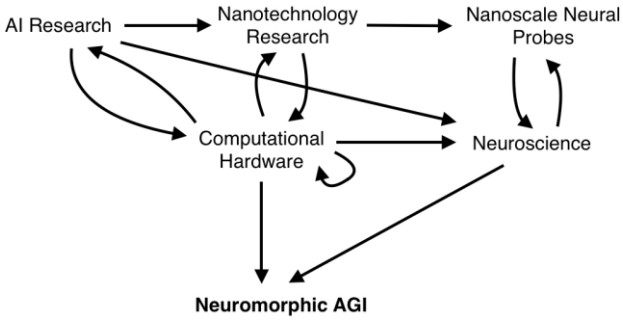

Figure 3: Technological pathway to Neuromorphic AGI. Neuroscience and computational hardware are the main technological requirements.

3.3 Whole Brain Emulation

The technological landscape of WBE looks quite similar to that of NAGI. Like NAGI, WBE would require computational hardware and neuroscientific knowledge. It should be noted, however, that the specific hardware and neuroscience requirements for WBE and NAGI may differ. WBE presumably requires more hardware, since WBE would be forced to simulate many details of brain function. In terms of neuroscience, NAGI would require greater conceptual understanding, while WBE would require greater understanding of details.

Another major difference between WBE and NAGI is the role of nanoscale neural probes. As we have argued elsewhere, it is unlikely WBE will be achieved without nanoscale neural probes [12]. In order to create a model of the human brain with enough fidelity to allow for WBE, we will arguably need the ability to study interactions within synapses, *in vivo*, and at large scale. Clearly, destructive brain scanning techniques (such as scanning electron microscopy) cannot achieve this alone, as these techniques destroy the brain and can therefore only be used to determine structure – not brain activity. Current large scale, nondestructive brain scanning techniques (such as MRI) don't appear capable of fulfilling this task either, as their resolution is too limited (and will likely run into harder limits such as any imposed by the skull). Single cell techniques (such as the patch clamp) can monitor single neurons or small groups of neurons for a few signals, but not large circuits of neurons for many types of chemicals. On the other hand, yet to be developed nanoscale neural probes would be able to fulfill all these tasks.

It should be noted that Sandberg and Bostrom have argued that WBE could be achieved without nanotechnology. They propose that brain architecture could be determined by automation of destructive scanning techniques similar to those that exists today (such as electron microscopy), using many of such automated machines in parallel. In order to model neuronal activity to the necessary precision, they suggest using a combination of this large-scale scanning and wet experiments (such as *in vitro* experiments on neurons) to create models, which can then be analyzed and used to guide further experiments, until the model is sufficiently refined [13]. We are personally very doubtful that such a scheme would allow for gathering the neuroscientific detail necessary for WBE. It is a well-known understatement to say that the brain's method of information processing is complicated, but what's not well appreciated is that this complexity in information processing doesn't just apply to how neurons are arranged – it includes many subcellular processes. Historically, scientists thought of the brain as consisting of a bunch of neuronal nodes that pass information simply in the form of "spikes" across passive synapses, yet we now know that reality is not so simple. In reality, neurons aren't simply nodes, but instead there is a large diversity of neuron types (with different behaviors), and computation is performed within the neuron cell bodies, within the axons, and within the dendrites [14]. Neurons don't communicate just through electrical signals either – around 10 common neurotransmitters and 200 uncommon neuromodulators are implicated in neuronal interaction, and neurons can even communicate without direct synaptic communication, such as via ephaptic coupling (nerve fiber coupling via local electric fields) and via chemical diffusion in the extracellular space [15][16]. Synapses themselves show a large diversity of types, and far from being simply conveyers of information, they play an active role in information processing [17]. In addition to neurons, glial cells (brain cells that outnumber neurons 10:1 but that have historically been ignored since they do not communicate via electrical impulses) can influence neurotransmission [18]. These findings have largely come as surprises to the scientific community, and it would be naïve to assume that we won't find any more surprises. The overall picture of the brain that we get from considering these findings is that much information processing occurs in the brain on the subcellular scale, which happens to be the nanoscale. In order to understand this information processing well enough to perform WBE, it is likely that nanotechnology will be essential. In biology, systems can act quite different *in vitro* versus *in vivo*. For a system as complicated as the brain, this would be particularly expected. Therefore, not only would brain activity likely need to be studied on the nanoscale, but it would likely need to be studied on the nanoscale *in vivo*. Nanoscale neural probes are the only foreseeable technology that could yield such information. Incidentally, several ideas for using nanotechnology to map the activity in the brain have been proposed, and the general idea of using nanotechnology to map activity in the brain is central to

the United State's multi-year, multi-billion dollar BRAIN Initiative [19][20].

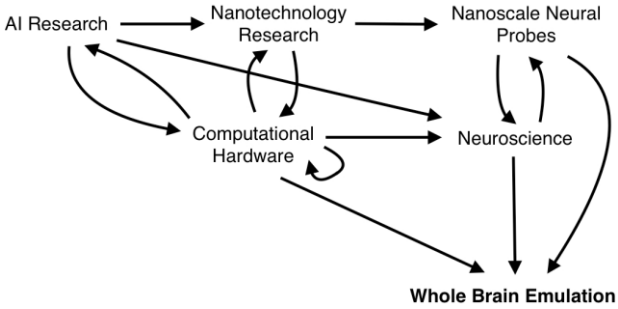

Figure 4: Technological pathway to Whole Brain Emulation. Computational hardware, neuroscience, and nanoscale neural probes are the major technological requirements. Note that the biggest difference between the pathway to Neuromorphic AGI and Whole Brain Emulation is the role of nanoscale neural probes in WBE.

While nanoscale neural probes are likely necessary for WBE, they are less likely to be necessary for NAGI. Since NAGI wouldn't need a one-to-one mapping with a particular brain, the specific details about the roles of each of these subcellular parts may not be needed – if certain general principles of how the brain processes information can be found via methods other than nanoscale neural probes, that might be good enough for NAGI. Even if there are gaps in the understanding of how the brain operates, it may be possible to create NAGI without understanding those gaps, by instead developing other algorithms that process information in a manner different from exactly how the brain does, yet that still accomplish the same general tasks.

3.4 The larger picture

Putting the aforementioned trends into a larger technological landscape yields a complicated and interconnected picture. Since our reasoning has included several simplifications (such as breaking AGI into only three distinct types), the real picture is undoubtedly more complicated. Accordingly, we must recognize that most implications are uncertain and open to revision upon further analysis. Having said that, we believe several implications can be produced.

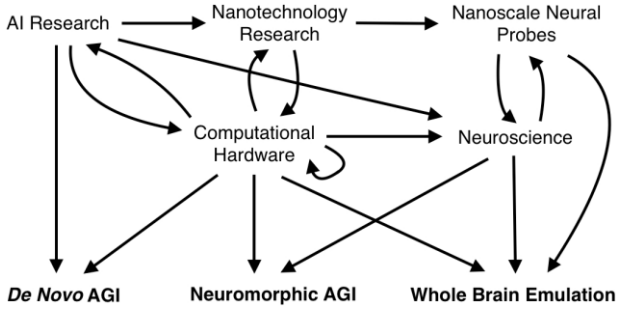

Figure 5: The technological landscape for the three main paths to AGI.

Computational hardware is the only major technological trend directly required for all three types of AGI, and it has a complicated relationship with AI safety. Larger computational hardware does have the potential benefit of favoring WBE over NAGI (compared to situations where smaller hardware may allow for NAGI but not WBE). On the other hand, there are a couple possible problems associated with computational hardware getting *too* large. First, such a situation could allow for a "hardware overhang" – where the massive computational resources available at the introduction of AGI enable early AGI to be particularly powerful or large in number. This situation may be more disruptive than if AGI were created in a context without such an overhang. Second, larger computational resources might allow for more dangerous methods of creating AGI. For example, it may be possible to "brute-force" an AGI without understanding it well. Alternatively, such large resources may allow for creating AGI through genetic algorithms. Since the use of genetic algorithms can lead to surprising results, even if the starting inputs for AI were relatively well aligned with human values, mutation and Darwinian pressures could cause the values to drift considerably astray.

AI research and neuroscience research similarly hold vague positions regarding safety. The more one favors dnAGI over WBE/NAGI (either by thinking WBE isn't much safer than dnAGI, or that if we pursue WBE/NAGI we will probably end up with NAGI) the more one should support AI research above neuroscience research (and the other way around if you disfavor dnAGI).

Development of both of these technologies requires caveats, however. For dnAGI, it is important that so-called AI safety research keeps pace with AI research, such that dnAGI isn't developed before we can align it with human values. For neuroscience, it may be bad if nanoscale neural probes lag significantly behind the neuroscience, as that state of affairs favors NAGI over WBE.

For each of these areas, it is more prudent to focus on accelerating the safeguard technology (AI safety research and nanoscale neural probes) instead of attempting to slow down the technology that poses a risk. For both of these areas, there are many more people working on the risky technology than on the safeguard technology, meaning that an individual can likely have a proportionately larger impact on development of the safeguard technology. Vested interests additionally imply that slowing down the risky technologies would be quite difficult. Advocating for the slowing of progress on such technologies could additionally cause a backlash, leading to actors in the field to dismiss all calls for safety as alarmist or neo-Luddite. Furthermore, even if calls for slowing down the risky technologies led certain safety conscious actors to refrain from pursuing such technologies, that would mean those left to pursue them would be less safety conscious (and in a worst case scenario, development of such technologies could be pushed underground).

Nanoscale neural probes require a bit of an elaboration, since their impact is more nuanced. Not only would they favor WBE over NAGI, but they also would favor NAGI over dnAGI. Even if one is generally skeptical of the WBE/NAGI pathway, these probes still

might provide promise. Since NAGI and WBE share many characteristics, it is reasonable to assume that they might be developed relatively close in time to each other, ignoring the impact that one form of AGI would have on creating another form. Since these two technologies are quite different from dnAGI, it is also reasonable to assume a greater chance that NAGI and dnAGI would be developed further apart from each other in time (again, ignoring the fact that one form of AGI could help create other forms). Therefore, even if advancements in nanoscale neural probes not only accelerate WBE as compared to NAGI by a certain amount of time (say, years), but also accelerate NAGI over dnAGI by a similar amount of time, the net effect may still be a decrease in the likelihood of NAGI. This is because if NAGI and WBE are in a closer race, it is more likely to tip the scale from NAGI to WBE than it is to tip the scale from dnAGI to NAGI in a less close race.

Nanoscale neural probes may further provide a particularly powerful means of human augmentation. Such augmentation wouldn't on its face favor any particular technological trend over any other one, but may well increase general human competence and decrease the chance of making a mistake in AI alignment. Furthermore, such probes may enable improved brain computer interfaces. This could aid proposals for making AI safe by having the AI act as an extension of humans.

4 Conclusion

In this piece, we examined safety concerns around the three major proposals for AGI: dnAGI, NAGI, and WBE. NAGI likely would be the least safe due to messiness and lack of other safeguards, and the inherent human safeguards in WBE would likely make it the safest. Even though with WBE a second transition to dnAGI would likely subsequently take place, we argued that a WBE-first path is still preferable (assuming such a path advanced WBE over NAGI to an extent that the chances of getting NAGI first did not increase), since WBE could aid a safer transition to dnAGI.

We also examined the major technological trends leading to these three types of AGI. While explicit AI safety research has been accepted as a means to ensure dnAGI is safe, we found that another path to increasing AI safety could be provided by the development of nanoscale neural probes, which would favor WBE over NAGI.

Of all the trends we've examined, nanotechnology research, and relatedly nanoscale neural probes, has traditionally been the most neglected by the AI safety community. This makes sense, given the fact that nanotechnology research is further removed from AI research than any of the other trends listed, and because it isn't obviously related to AI, in contrast with AI research, hardware, and neuroscience.

Since nanotechnology holds many implications for AGI, further research should not ignore the implications that nanotechnology may hold.

5 References

[1] Müller, V. C. and Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. *Fundamental Issues of Artificial Intelligence*, Synthese Library; Berlin: Springer, 553-571.

[2] Hawking, S. *et al*. (2014). Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough?' *The Independent*.

[3] Farquhar, S. (2017). Changes in funding in the AI safety field. *The Center for Effective Altruism*. https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field/

[4] Amodei, D., Olah, C. *et al*. (2016). Concrete Problems in AI Safety. *ArXiv:1606.06565 [cs.AI]*.

[5] Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *ArXiv:cs/0004001 [cs.AI]*

[6] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[7] Yudkowsky, El. (2004). Coherent Extrapolated Volition. *The Singularity Institute*, San Francisco, CA. https://intelligence.org/files/CEV.pdf

[8] Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

[9] Amodei, D. Interviewed by Wiblin, R. (2017). Podcast: How to train for a job developing AI at OpenAI or DeepMind. *80,000 Hours.* https://80000hours.org/2017/07/podcast-the-world-needs-ai-researchers-heres-how-to-become-one/

[10] Urban, T. (2017). Neuralink and the Brain's Magical Future. *Wait But Why*. https://waitbutwhy.com/2017/04/neuralink.html

[11] Shulman, C. and Sandberg, A. (2010). Implications of a Software-Limited Singularity. *Machine Intelligence Research Institute*, Berkeley, CA. https://intelligence.org/files/SoftwareLimited.pdf

[12] Eth, D. *et al*. (2013). The prospects of whole brain emulation within the next half-century. *Journal of Artificial General Intelligence*, 4(3), 130-152.

[13] Sandberg, A. and Bostrom, N. (2008). Whole Brain Emulation: A Roadmap. *Future of Humanity Institute*, Oxford University, Technical Report #2008-3.

[14] Sidiropoulou, K. *et al*. (2006). Inside the brain of a neuron. *EMBO Reports*, 7(9) ,886-892.

[15] Anastassious, C. A., Perin, R. *et al*. (2011). Ephaptic coupling of cortical neurons. *Nature Neuroscience*, 14(2), 217-223.

[16] Vizi, ES, *et al*. (2010). Non-synaptic receptors and transporters involved in brain functions and targets of drug treatment. *British Journal of Pharmacology*, 160, 785-809.

[17] O'Rourke, N. A. *et al*. (2012). Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nature Reviews: Neuroscience*, 13, 365-379.

[18] Fields, R. D. *et al*. (2014). Glial Biology in Learning and Cognition. *The Neuroscientist*, 20(5), 426-431.
[19] Marblestone, A. H., Zamft, B. M. *et al*. (2013). Physical principles for scalable neural recording. *Frontiers in Computational Neuroscience*, 7(137), 1-34.
[20] Editorial. (2014). Brain activity. *Nature Nanotechnology*, 9, 85.