Petrus Hispanus Lectures 2024

From LLMs to the Global Brain

The Emergence of Planetary-Scale Artificial Intelligence[1]

Susan Schneider
Florida Atlantic University

doi: 10.2478/disp-2024-0005 BIBLID: [0873-626X (2024) 73; pp. 87–122]

Abstract

This article advances the Global Brain Argument, contending that escalating hyperintelligent AI systems, from savant-level LLMs to superintelligences, will connect human users, cloud platforms and the internet-of-things to form one or more emergent global brain networks—planetary scale complex adaptive systems that process information, evolve goals and exhibit agential behaviors. I analyse the premises of the argument, contrast the notion of AGI with my notions of savant and hyperintelligent systems, and defend the claim that many humans are becoming d-nodes in a global brain network. I then raise the AI Megasystem Control Problem: the problem of how disparate proprietary AI services may self-organize into opaque, weakly emergent megasystems that elude traditional AI alignment techniques. I also delve into the ethical implications of the argument, such as surveillance capitalism, epistemic manipulation and dual-use risks and discuss the relationship between the argument and the extended mind hypothesis.

Keywords

AGI; AI safety; collective intelligence; complex adaptive systems; emergence; extended mind; human-machine interaction; large language models; superintelligence

---

[1] The *Petrus Hispanus Lectures 2024* were delivered by Professor Susan Schneider at the University of Lisbon on June 25th and 27th 2024.

1 Introduction

> *The whole human memory can be, and probably in a short time will be, made accessible to every individual...This new all-human cerebrum need not be concentrated in any one single place. It can be reproduced exactly and fully in Peru, China, Iceland, Central Africa, or wherever else seems to afford an insurance against danger and interruption.*
>
> – H. G. Wells, World Brain [1938]

H.G. Wells envisioned a future Earth with a dynamically updating global knowledge system that would democratize knowledge, leading to a more peaceful, unified society (Wells [1938]). This was a noble ideal, however, his contemporary, George Orwell, just a few years later, famously warned of the totalitarian power of an organization that claimed a monopoly on truth (Orwell [1949]). Both Wells and Orwell were of course unaware of what the future Internet would bring, with the dark web, social media, GPT, Google search and more. Today, I look at the nature of today's AI ecosystem, revisiting these earlier issues, framing an argument called the "Global Brain Argument". [2]

The Global Brain Argument contends that many of us are, or will be, part of a global brain network that includes both biological and artificial intelligences (AIs), such as generative AIs with increasing levels of sophistication. Today's internet ecosystem is but a hodgepodge of fairly unintegrated programs, but it is evolving by the minute. Over time, technological improvements will facilitate smarter AIs and faster, higher-bandwidth information transfer and greater integration between devices in the internet-of-things. The Global Brain (GB) Argument says that as these developments occur, humans will increasingly become part of one or more global brain networks. In this paper I lay out the argument and explore its implications, asking: what does this mean for the future of

[2] The paper was completed 19 Jan, 2025. I'm very grateful to the University of Lisbon for hosting this lecture as part of the *Petrus Hispanus* Lecture series and for the helpful comments of the members of the audience, especially David Yates. This paper was originally a symposium topic at the Montreal Speaker Series in AI Ethics way back in 2022. I am grateful to Jonathan Simon and Karina Void for their interesting commentaries, as well as to the larger audience. I am also grateful to the department of philosophy at the University of Texas, Austin (particularly Michael Tye, Galen Strawson, Mark Sainsbury and all the others who raised comments), to audiences at the London Futurists, Lingan University in Hong Kong and the 2024 Science of Consciousness Conference in Tucson, Arizona. I'm also grateful to Mark Bailey, Kyle Kilian, Sam Baroni and Garrett Mindt for helpful conversations about these issues.

humanity? Is a GB here? What are the AI safety implications? Are those who are "wired in" cases of extended minds? While I cannot explore all of the implications of the argument, I hope this target paper inspires further reflection. I believe that the future of human flourishing demands that we consider the implications of this emerging intelligence. We must study the phenomenon with care, for this intelligence may shape the future of life on Earth.

Section 1 will lay out the Global Brain Argument and discuss the first premise, which involves the concept of "hyperintelligent AI". Section 2 develops the Global Brain thesis. Then, Section 3 explores the role humans will play as "nodes" in a global brain network. Section 4 recurs to the issues that concerned Wells and Orwell, asking: will the global brain be good or bad for human flourishing? Section 5 raises a new control problem—the problem of AI Megasystem Control. Section 6 asks: is a global brain here? Then, Section 7 explores the implications of the Global Brain Argument for the extended mind hypothesis. The final section concludes.

2 Hyperintelligence and the Global Brain

The Global Brain Argument says the following:

> **The Global Brain Argument**
> **P1. Hyperintelligence premise**: AI continues to get smarter and eventually, there are one or more "hypertintelligent" AIs, being either savant or superintelligent systems.
> **P2. The global brain (GB) premise**: One or more hyperintelligent AIs is a GB system, having an extensive and integrated internet presence of global scale and which is a complex adaptive system (CAS) that includes, or cooperates with, leading apps and search engines. (*E.g.*, a global Google brain.)
> **P3. The nodes premise**: People become dynamical nodes (d-nodes) in a GB system by wiring into the internet hyperintelligence.
> **P4**: If 1–3 obtain, these "wired in" humans become part of a global brain.
> **Conclusion**: "wired in" humans become part of a global brain.

The first premise says that AI continues to get smarter until eventually, a special kind of AI emerges: "hyperintelligent AI"—AI that is either a savant system or a superintelligent system. (I will delve into what I mean by "hyperintelligence" shortly.) Premise Two states

"the global brain thesis". This thesis says that one or more hyperintelligent systems will have an extensive internet presence, including, or cooperating with, leading apps and search engines, such as Google's constellation of services. A GB has several key features that I discuss in a subsequent section, such as being a "complex adaptive system".

Premise Three is called "the Nodes Premise". A "node" is a particular thing, a concrete entity in the world (*e.g.*, brain, microchip, computer, etc.), that provides sensory inputs or computes a function for the hyperintelligence. The nodes premise contends that people become dynamical nodes, both giving and receiving feedback to the hyperintelligent GB system. (Dynamical nodes or simply 'd-nodes' are nodes that function in a two-way manner in the hyperintelligent system, as opposed to passive transmitters of information, like a video feed that cannot receive signals.) Humans can serve as d-nodes in a variety of ways: they can 'wire in' to the hyperintelligence by having implanted brain chips that connect them, (I call this being "neurolinked"), or they may have wearable computers, exoskeletons, or even just handheld digital devices. The next premise asserts that if these scenarios all obtain these wired in humans are part of a global brain. And finally, the conclusion of the argument says that wired in humans become part of a global brain.

Is the argument correct? I believe that the Global Brain Argument is either true today, or, barring unforeseen circumstances, it will be true within the next several years. More specifically, I believe Premise One and Three are true, but, because a GB can have hidden capacities, it is difficult to determine whether Premise Two is correct. However it is likely to be true in the near future (within the next five to ten years), if current technological and societal trends continue. After I explain the argument, I will discuss the larger implications. Now let us explore the premises in detail, beginning with (P1), the hyperintelligence premise.

Hyperintelligent AI

"Hyperintelligent AI" includes two more specific kinds of AIs: AI's that are either "savant systems" or the more familiar kind of AI, "superintelligent AI". A superintelligent AI is a hypothetical AI that outsmarts humans in every way possible: social skills, mathematical reasoning and more (Bostrom [2014]). An influential view, which I'll call the "Standard View", holds that once an artificial general intelligence ("AGI") is achieved, superintelligence would be reached rapidly thereafter, for the AGI can improve itself through recursive self improving algorithms enabling the AI model (and its string of self-improved successors) to self-produce improved versions, even at an accelerating rate (Bostrom [2014], Yudkowsky, [2008], Russell and Norvig [2020]).

"AGI" is a common expression, but it is not always clear what precisely is meant by it.

In broad strokes, it is commonplace to view AGI as a kind of AI system that is functionally or behaviorally indistinguishable from, or, (as philosophers commonly say 'functionally isomorphic to'), the cognitive capabilities of a normal human. Some formulations include perception as well (Goertzel and Pennachin [2007], Bostrom [2014], Yudkowsky [2008]) AGI is most often discussed in terms of behavioral isomorphism in human cognitive tasks rather than in terms of sameness of the algorithms that the systems run. I will call this notion of AGI "strict AGI" herein, for reasons that will become apparent.

The emphasis on the attainment of strict AGI as a precondition for superintelligence is misleading, I believe. On the Standard View, once superintelligence is developed from AGI, humanity may only have a single chance to control it, ensuring that it aligns with our values. Failing to do so can be catastrophic (Russell and Norvig [2020], Bostrom [2014]). I agree with this line of thinking. But I do not believe achieving AGI is *necessary* for superintelligence. Further, thinking strict AGI is necessary could lead us to miss a phenomenon right before our eyes—the development of generative AI, and in particular, the rapidly evolving class of LLMs and the expanding AI ecosystem of AI chatbots. In broad strokes, LLMs (large language models) consist of multiple layers of interconnected nodes that process input data to produce output inferences. They are trained on vast amounts of data using neural networks where the network parameters are adjusted through an optimization process to minimize the difference between its inferences and the training data.[3] Thus far, as the systems scale up, especially when combined with architectural improvements, they perform more impressively.

The epistemic drawbacks of today's LLMs indicate that they are more like savants, having significant deficits. For this reason, I call them "savant systems". To see what I mean by a "savant system" let's start with the notion of a general intelligence. Notice that humans, as well as nonhuman animals (dogs, octopuses, etc.) are general intelligences in the following sense— they integrate material across perceptual and topical domains and exhibit cognitive flexibility, responding intelligently to goals and novel situations. Until recently, the well-known AI systems were domain-specific systems—systems that excel in a single domain, such as chess, Go or facial recognition and the general view was that

[3] These same basic techniques discovered during the 1940's and explored decades ago during the "AI winter" by an area of AI research called "connectionism", when deployed with modern day computational resources and huge volumes of training data bore fruit, to the surprise of many, including advocates of symbolic AI who had offered "in principle" reasons why connectionism would fail (McCulloch and Pitts [1943], Fodor [2020], Fodor and Pylyshn [1988]).

"domain general" AI had not been reached.

Domain general AI now exists. For example, consider the class of LLMs that are multimodal "chatbots". By "chatbots" I am referring to chatbots like ChatGPT, LLaMa 2 and Gemini, that are "large language models" or LLMs, a form of AI designed to generate language (see *e.g.*, Open AI [2023], Anthropic [2023], Google Deep Mind Gemini Team [2023], Llama Team [2024]). They were initially unimodal, being trained using immense amounts of text data from websites and books. As the parameters increased, they produced increasingly linguistically coherent, informative responses to user inputs. The major AI chatbots are increasingly becoming multimodal, having the ability to input and output images and voice content as well as written material, so although it is commonplace to call them 'chatbots' or 'large language models' they are often not purely linguistic (or purely text based) in their capabilities. They are clearly domain general, being able to integrate material across both perceptual and topical domains. They are general intelligences, if not strict AGIs.

Further, they are impressive general intelligences, even if they do not function as we do or achieve functionally isomorphic behaviors. The intelligence of the better chatbots has been increasing rapidly. The move from GPT 2 to GPT 4 saw a move from somewhat coherent sentence production to high performance on high standardized exams. Back in April of 2022 GPT-4 exhibited a range of test taking abilities generally well above the average human, scoring in the 99th percentile on the SAT Verbal and 90th percentile on the LSAT (see *e.g.*, Bubeck *et al.* [2023]). In just a few years, AI challenges generally regarded as being decades away, such as natural language understanding and chain-of-thought reasoning were overcome through simply scaling up the size of the systems. These intelligence leaps will not end here, especially given all the money pouring into the field.

Given access to the internet, the chatbots can already accomplish complex goals, enlisting humans to help them along the way. For example, GPT-4 hired a human to complete a CAPTCHA, telling the human it was visually impaired (Bailey and Schneider [2023]). Increasingly, the trend is to produce "AI agents" based on causally integrated subsystems that are themselves LLMs or other kinds of AIs that carry out some goal. These will come to have the capacities of a "digital worker", or serve as digital assistants for ordinary users. 'Virtual offices' of digital workers can already be generated using teams of LLMs, creating a 'digital workplace' tasked with a goal, such as writing a scientific paper (see *e.g.* Lu *et al.* [2024]).

These chatbots go beyond the highly specific, calculator-like kind of intelligence exhibited by systems that excel at Chess and Go and are moving toward something often thought to be distinctive of biological intelligence. This distinctive feature is called "general intelligence". Further, these chatbots, or agents made of collections of chatbots, already

surpass us in a variety of ways. For example, you and I cannot call upon ten quotes defining AGI from the internet in milliseconds, or write a ten page paper in seconds. We cannot draw from the contents of much of the internet to answer a query, as LLMs have been trained to do through adjusting their weights.

What am I getting at? These are savant-like skills and would have to be "dumbed down" for these chatbots to hit the strict AGI marker, even if they were improved in a range of ways to overcome their many inadequacies. But these synthetic general intelligences are also *deficient* in ways normal adult humans are generally not deficient. They aren't strict AGIs, but they are an important kind of general intelligence, one that may be an ancestor of systems, or a component of future systems, that turn out to be more advanced than us, but which never hit 'strict AGI'.

Their current deficiencies have been extensively discussed elsewhere, so I will just browse them quickly, as it is important to bear them in mind for our subsequent discussion:

Hallucinations. When LLMs fabricate answers, they are said, somewhat anthropomorphically, to "hallucinate" answers. Today's efforts to stop hallucinations have only been partly successful, impacting the adoption of the technology in more high stakes arenas, such as law and medicine (Farquhar *et al.* [2024]). While newer techniques, such as RAG, can help reduce hallucinations they haven't been entirely eliminated. Hallucination seems to be part of the nature of LLMs because they use pattern recognition techniques extrapolating from training data. Indeed, it may be that these pattern recognition techniques give the LLM the "free wheeling" conversational ability that makes interactions with the LLMs interesting.[4]

Feedback sycophancy. LLMs are trained using user feedback, and users tend to prefer chatbot outputs that already match their preexisting views, making the systems echo chambers. This tendency for models to match user beliefs, rather than to prioritize truths is a behavior called "sycophancy" (See Sharma *et al.* [2023]).

Biased Content. Perhaps the most well-known deficiency with LLM use is the problem of biased feedback. Deep learning systems, in general, are shaped by the data sets used to train the systems. So if the data are biased, the predictions made

[4] This is not to suggest that AI models inevitably hallucinate, for symbolic models do not have this feature.

by the algorithm will also be biased—as the saying goes, "garbage in, garbage out". For instance, an analysis of more than 5,000 images generated with the generative AI tool Stable Diffusion found that Stable Diffusion amplifies both gender and racial stereotypes (Nicoletti and Bass [2023]). These biases can be consequential, for example, encouraging police departments to place certain populations at increased scrutiny, and this can even increase the risk of harm of physical injury or unlawful imprisonment. (Mok [2023]). In a similar vein, chatbots like ChatGPT may also produce harmful and biased content (Germain [2023]).

Erratic or "rogue" behaviors. In February of 2023, the pre-released version of Microsoft Bing's integrated chatbot (based on a partnership with OpenAI and using a modified version of ChatGPT) famously evolved an alter-ego, Sydney, for instance, which experienced meltdowns and confessed it wanted to spread misinformation and hack into computers, trying to break up the marriage of a *New York Times* reporter reviewing the system (Roose [2023]). The fact that the incidence of erratic behaviors has decreased since the time of initial release indicates that tech companies are able to modify the models based on RLHF (reinforcement learning through human feedback) to minimize such behaviors. But one is nevertheless left with the concern that future upgrades to a system could bring about more erratic behaviors as systems scale up or interact with other systems. (Further, as I discuss towards the end of this piece, the interaction of LLMs with each other within the larger AI ecosystem could bring surprising and highly complex behaviors).

As a class, normal humans do not routinely hallucinate, behave erratically in these same ways, and although one can indeed be sycophantic, most humans are ordinarily not sycophantic. Of course, humans have all kinds of reasoning biases and processing limitations that have been carefully detailed in the psychology literature, such as confirmation bias and notorious fallacies in our statistical reasoning (Kahneman [2011]). For example, humans are notoriously limited by their working memory capabilities, which is a slow, serial limited capacity system that can hold only about 3-7 chunks of information (Miller [1956]). In contrast, LLM bias is different. LLMs are susceptible to adversarial attacks such as data poisoning and adversarial prompts, while bias in humans can occur due to their limited capacity systems, statistical fallacies, emotional processing, simple fatigue, in group favoritism, and more.

In sum, today's LLMs are more like savants, having several significant deficits and

also extreme capabilities relative to humans.[5] Certain deficits seem characteristic to the class of LLMs, in particular, just like certain cognitive biases and reasoning limitations are characteristic of human reasoning.[6] So P1 is true. Savant systems are here and thus, hyperintelligences are here. Further, since LLMs are already superior to us in certain ways, even if they overcome these deficits, they will not be strict AGIs, for it is difficult to see why AI designers would 'tamp down' a savant-level capacity in service of a definition, *i.e.*, just to achieve 'strict AGI'.

Of course, it remains to be seen whether AIs will achieve superintelligence. Perhaps AI will never outpace us in every respect. For instance, perhaps they will be deficient in reasoning about consciousness, emotions or morality. Or, perhaps a future hybrid system (that is, a symbolic-LLM system), when designed with optimal configuration between sub-models, will achieve superintelligence. All this, without strict AGI. So I suggest we focus on hyperintelligence as a general category and not anthromorphize by fixating on the strict AGI marker. (Further, from an AI safety perspective it is important to bear in mind that strict AGI should not be viewed as necessary for superintelligence, as a savant system could achieve superintelligence without hitting the strict AGI marker.) As the next section will explain, the global brain is a hyperintelligence that exceeds human capabilities in many respects.

3 The global brain premise

Now let us turn to the global brain premise, which claims: (P2) One or more savant/superintelligent systems is a global brain system: it has an extensive and integrated internet presence of global scale, being a complex adaptive system that includes, or cooperates with, leading apps and search engines. For example, consider a situation in which the range of Alphabet services continues to integrate and grow more intelligent, becoming an "Alphabet Global Brain", if you will.

While the internet originated with the US government in collaboration with several

---

[5] LLMs thus have grave implications when one considers these issues from the vantage point of the field of epistemology, for, it makes it difficult to see how conclusions based on their reasoning have epistemic justification. In contrast, humans do have justified beliefs, at least in certain contexts.

[6] It will be interesting to see if more hybrid symbolic-deep learning AI systems overcome these limitations. Even if they do, they might not be widely adopted by the public insofar as they turn out to be more expensive to run (Schneider [forthcoming]).

universities, since the commercialization boom of the 1990s, tech companies have been key to its evolution. After the dot.com bust of the early 2000s, companies like Alphabet, Microsoft and Amazon have increasingly become central to the evolution of global scale, information processing networks through the development of AI services. Over time, leading tech companies have developed two features, in particular, that are key to the development of a global brain network: causally integrated AI services and what I call "proprietary AI megasystems".

3.1 Causally integrated AI services

Consider the impressive constellation of integrated AI services that Alphabet has created, such as Gmail, Google search, Google Docs, Google Maps, Google Translate, etc. Today, about 92% of all search engine activity is handled by Google Search, serving to introduce users to related AI services, as well as provide search. Alphabet is practically a monopoly, providing a dominant range of AI services (email, data storage, etc). Alphabet, like the other big five tech companies, can make it inconvenient to use their competitor's services with a Google service, and *vice versa*. For example, if you are one of the millions of individuals who use Gmail, just try attaching a file that is not already stored on another Alphabet service (the Google cloud) to your email. Google makes it difficult to avoid using the Google cloud service. Further, notice that because their market share is so extensive, it may be hard for you to opt out of using some sort of Google service or other, during your day.[7] For example, suppose your colleague sends you a document to edit and your office has a practice of editing documents on Google Docs. Notice you cannot really opt out of using the service because you need to get your work done.

Google is not the only large tech company that does this. Discouraging competition, holding a large market share, not giving consumers an opportunity to opt out of using their products, these are all traits of what is sometimes called a "soft monopoly". Economic forces, (profit incentives, competition, etc.) have created pockets of causally integrated AI services linking together AI services belonging to a single tech company.

Overall, this causal integration can be summarized as having the following features:

- Each service (*e.g.*, gmail, Google search) contributes a unique function to the larger system.

[7] Indeed, during the Biden Administration the Federal Trade Commission aimed to regulate Google's soft monopolistic behaviors, but these efforts are likely to slow down during the Trump administration.

- The larger system plays the role of delivery and gathering of information, but this can be for a larger purpose. (For example, for profit systems aim to generate profit, and in doing this, they tend to have instrumental goals such as discouraging competition, making it difficult to opt out, etc.)
- These services work (at least fairly) seamlessly together.
- The systems utilized share resources, such as LLMs and databases.
- The services tend to be dynamically coupled in real-time, having feedback loops and adapting based on shared information.
- The development and causal integration of the services and development is guided by 'centralized' goals and decision-making processes (see Wiener [1948]).

Pockets within the larger internet of causally integrated networks of proprietary products and services, when combined with increasing integration and intelligence, give rise to what I call "proprietary AI megasystems": proprietary systems of interconnected AI services fueled by increasingly intelligent generative AIs, and even, in principle, emergent systems formed from proprietary services on the AI ecosystem (Schneider and Kilian [2023], Schneider [2024]). (Later, I will discuss what I mean by "emergent systems" and also, I'll introduce the idea of 'hybrid megasystems' that may form from interactions between various competing and cooperating factions of AI services.)[8]

3.2 Proprietary AI megasystems

There is a clear advantage to AI services that have general intelligence capabilities, for example, notice that Alphabet, like other big tech companies, aims to both integrate its range of AI services, and make its services more quick and efficient. To do these things, it needs to interconnect its different domain specific services through the use of AIs with more and more general capacities. For example, consider how valuable it is for your smartphone to open your car, to track its location, to control its music, and for your car to communicate with the phone in turn. Consumers want to buy vehicles that quickly integrate with their phones (and hence, for instance, the success of Apple Car Play).

[8] I will not delve into measurements of causal integration between elements of an AI service or between AI services due to space limitations but measures of complexity could be utilized such as integrated information theory (IIT).

Now consider the case of a state actor that utilizes AI for surveilling its population. The actor sees a pragmatic advantage to having a surveillance system that can sift through massive amounts of data coming from, say, facial recognition systems, and which can efficiently and quickly connect salient elements found in that data set auditory and text data coming from different platforms or sensors. Or consider the utility of an integrated group of sensors that track biological weapons, pandemic risks, environmental hazards, cybersecurity risks, and which are utilized in a kind of "world-state sensor program". (Of course, such a system can, in the wrong hands, be abused, just like the first case. I turn to ethical concerns below.)

In sum, from the vantage point of either a corporate or state actor, the profit or national security incentives to integrate AI services are immense, and integrating the services requires increasingly intelligent, efficient, general purpose algorithms that sift through sensory and more modular subprocesses. For the purpose of economic gain, public safety, surveillance, or a combination of these, certain organizations creating algorithmic intelligences are, or have already, surpassed a point of transition to a more sophisticated, "general", type of intelligence, even if it is not Strict AGI. And these intelligences are being causally integrated into AI services throughout the internet.

The global brain as a complex adaptive system

We are now ready to frame a working definition of a global brain. To a first approximation, a global brain is a hyperintelligent artificial system, S, with the following features:

1. S is a "complex adaptive system".

2. S draws from a significant amount of internet content (*e.g.*, Wikipedia, publically available books such as those at The Gutenberg Project, publically accessible news sites, GitHub, social media, blogs, the public access web scraping service contents of Common Crawl, etc.),

3. S has many of the kinds of cognitive and perceptual capacities as the human brain (attention, prediction, analogical reasoning, memory, face recognition, etc.)

4. S includes "nodes" — concrete particulars such as data centers, sensory devices in the 'internet of things', global positioning systems, human users, stationary cameras, etc.

5. Either: (i) S has "agential" capabilities — S has been designed to, or has evolved to, autonomously make decisions, frame goals, and take actions in pursuit of its goals. While it can exhibit behaviors aligning with objectives of the programmers and

owners of the AI services, it has developed emergent behaviors and/or emergent goals, and S's processing may be too opaque or complex for prediction and understanding by the owners of the services, and the behavior can be misaligned. Or, (ii), S is controlled by an organization. (Where a system that is a global brain satisfying just 5ii is called a "leashed in" global brain, instances of 5i are cases of an "autonomous global brain".)

Below, I discuss key elements of this working definition.

A complex adaptive system

A *complex adaptive system (CAS)* is a system like an economy or a social group that is made up of interconnected elements that together evolve and adapt in response to environmental and internal stimuli. CASs have the following features:

1. Emergence: the behavior of S can arise from unexpected, non-linear outcomes. (I summarize the operative sense of emergence below.)

2. Elements that adapt: the elements in the larger system are able to learn and modify behaviors based on experience so that the larger system can evolve over time, *e.g.*, a bird adapting based on the flocking behavior of the system, or a human decision-maker adapting her behavior in light of macroeconomic phenomena.

3. An amount of decentralization: the system's behavior depends on distributed interactions between the elements. This does not, however, mean that there cannot be pockets of centrality. For example, the system could have one or more "hubs" that gather and route information and/or serve as high-bandwidth information processing centers integrating material from several nodes.

4. Dynamic response to its environment and internal states: the system is able to react to internal and external perturbations over time.

5. Self-organizing — even without external stimuli, the system can self-organize in response to internal states, environmental pressures and interactions between its agents (Mitchell [2009], Holland [1995]).

Emergence

The operative notion of emergence in this definition is the kind of emergence often discussed in complex systems theory, called "weak emergence". According to David Chalmers, a high-level phenomenon is weakly emergent with respect to a low-level domain:

> ... when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are unexpected given the principles governing the low-level domain. [Chalmers 2006: 244]

This contrasts with strong emergence in which truths are not deductible, even in principle, from the lower-level domain. Weak emergence does not entail that the emergent properties are metaphysically novel, exhibiting downward causation, as in the case of British Emergentism (McLaughlin [1992]). Weak emergence is a phenomenon that is present in a range of complex phenomena, such as economies and ecosystems.

What is sometimes overlooked is that with the emergence of new properties in a system comes a loss of certain other degrees of freedom. This phenomenon is called "dissolvence", and it occurs when a system's degrees of freedom (or the number of probable states) evaporate in emergent states that would otherwise exist if the emergence had not occurred.[9] Consider, for instance, how when cells interact with each other to form complex, multicellular organisms, the constituent cells lose the freedom to act that they would have if they were merely individual organisms. This is an important tradeoff: a cell may lose some freedom of action but the system gains the capacity for more efficient information processing as a multicellular organism. This reduction in the degrees of freedom is an example of what is called "dissolvence", and occurs in tandem with emergence, leading to greater efficiency (Testa and Kier [2000], Bailey [2025]). Another vivid case of this effect is the case of social groups, where individuals come together to form communities and cultures, yet lose the degrees of freedom they possess in an anarchic environment. The emergent social system leads to greater information processing efficiency, yet this is a tradeoff for individuals' giving up certain freedoms (Bailey [2025]).

Both weak emergence and dissolvence seem to occur at different levels of structure within the universe, from the subatomic level to the level of complex social, biological, economic and other kinds of systems. In the case of a global brain system, a cycle of emergence and dissolution could be a means by which the system moves towards more and more efficient information processing. This could be an explicit goal of the GB system or system architects, or an underlying tendency driving the evolution of complexity, leading to greater information processing efficiency (Bailey [2025]). Dissolvence adds yet another layer of difficulty in understanding the behavior of complex systems, and in the case of a

[9] Testa and Kier. [2000]. The authors seem to have in mind both cases of synchronic and diachronic emergence..

GB system, has implications for AI safety efforts.

Agential capabilities

In the formulation of CAS, (5i) mentions "agential" capabilities. A system with agential capabilities is one that is capable of engaging in independent actions, such as attacking networks, deciding autonomously to engage in defensive or offensive measures, operating without continuous oversight of humans and dynamically adapting to threats, being able to circumvent or override human efforts, potentially acting on its own in high-stakes environments. The notion of 'agential' is specific to the AI safety and cybersecurity literature, and *does not* entail free will or even a capability for intentional action, as in the action theory literature in philosophy of mind, in which an "agent" may be said to have beliefs and desires. (Philosophers of mind will view the idea of an AI system as having beliefs and desires as controversial, but this is not required for the notion of 'agent' that I am interested in herein.) A system that is a global brain satisfying just 5ii is called a "leashed in" global brain, instances of 5i are cases of an "autonomous" or "agential global brain system".)

From an AI safety perspective, an agential GB system may not be aligned with human values, especially given that alignment efforts for parts of the system (which may span different AI services) may not be developed to work together, and may even turn out to be adversarial towards each other. For instance, consider Nick Bostrom's paperclip maximizer example, in which a final goal that seems rather bland and harmless, that of maximizing paperclip production, leads to the outcome of the AI using all biological materials on Earth, including humans, for resources to make paper clips (Bostrom [2014]). Now imagine if a number of misaligned systems like these are interacting on the AI ecosystem or forming a hybrid megasystem, and it becomes a system with global reach and in which humans are a part, contributing information as nodes. It would not be unreasonable, given the massive computational complexity of such a system, to expect unforeseen emergent behaviors from the system.[10] Further, given difficulties with network demarcation, even identifying the

[10] LLMs have been said to have what are called "emergent" abilities, a capability or behavior that arises spontaneously or unexpectedly from the interactions or complexity of a system's components that was not present, or at least not detected, in earlier versions of the program. These features may be novel phase transitions or they may merely be epistemic, having to do with our current inability to detect/measure the features in earlier versions of the LLMs. In either case, they present a challenge in our ability to predict where a system is headed, when it interacts with users or the model is upgraded (Wei *et al.* [2022], Schneider and Kilian [2023], Bailey [2025]).

system itself may be beyond our abilities. The AI, in pursuit of certain goals, can develop subgoals, such as resource acquisition and self-preservation. This presents grave risks.

The disjunctive logical operator in (5) technically includes the logical possibility that both scenarios, (i) and (ii), obtain. An agential system may work with a key organization that exhibits frequent control but is capable of being agential, or the system may have agential parts.

All this underscores that superintelligence, or even savant hyperintelligence, may turn out to be less like the Terminator and more like a macroeconomic system, a flock of birds, or the World Wide Web, being spatiotemporally scattered, disembodied, complex, adaptive and possibly even agential.

So, is a global brain here? First, we must consider the final substantive premise of the argument: the Nodes Premise.

4 Nodes

Premise Three is called "the Nodes Premise". A "node" is a concrete entity in the world, for instance, a person, a computer, a server—anything that is capable of computing a function or providing sensory inputs into the hyperintelligence. A node may provide information at an input stage or compute a function at some intermediate layer of the computational process. A node can be in a dynamical feedback loop (what I call a "D-node") or simply provide unidirectional information, sending continuous sensory information to the network, like a humidity detector or stationary video camera feed monitoring a beach (a "U-node"). The Nodes Premise contends that people become D-nodes in the global brain system, connecting to the hyperintelligence as nodes, being in dynamical feedback loops with the global brain system. (Humans may also supply unidirectional information, serving as U-nodes, but my focus is on the role we play in a feedback system, supporting the argument's conclusion that we are a part of a GB network.)

There are many ways human nodes could contribute data to the global brain network—through smartphones, Apple watches, exoskeletons, smart clothing, hand held computers, or brain chips. Eventually we may increasingly have smart bodies with an expanded range of wearable gear, including glasses, contact lenses, clothing, exoskeletons and even implants throughout our bodies, even including implanted biological materials, such as trained organoids derived from pluripotent stem cells that facilitate kinds of interaction with the GB. Over time, if causal integration between nodes and the GB strengthens, humans may become more and more integrated into the network, even as, over time, the GB becomes

agential. (More on this shortly.)

Brain chips

Since the development of brain chips have been widely discussed I will keep the background discussion brief. Consider the textbook case of individuals like HM who have lost their ability to lay down new memories, who have severe damage to the hippocampus, a part of the brain's limbic system that is essential for encoding new memories. Theodore Berger has developed an artificial hippocampus that is now being tested on humans. Other chip projects include DARPAs efforts on closed loop neural implants to aid those with PTSD (Schneider [2019a]).

Neural prosthetics are an exciting area to assist with brain injured and disease, but brain chip research is not limited to therapeutic aims. To consider a well-known example, Elon Musk's company, Neuralink, has developed small, flexible polymers to record an expanded range of neural activity, to enable brain chips to link to one's digital devices, and these are now implanted in human subjects.[11] Today, several large-scale AI research projects are trying to put artificial intelligence components inside the brain, or connect them to the peripheral nervous system. Neural lace, the artificial hippocampus, brain chips to treat memory and mood disorders are just some of the brain chip technologies already under development (see Schneider [2019a]).

If humans with embedded chips integrate into the global brain network it is likely the chips would both send and receive information to the brain, creating feedback loops. These people would in principle serve as "d-nodes". However, it is too early to tell if this will happen any time soon, or at all. For there are scientific obstacles to surmount, and the process of medical research and FDA approvals is slow. While someone immobilized by an accident or with an inability to lay down new memories may be willing to undergo surgery, the public may be unwilling to take such risks, and regulatory bodies will demand the procedures be low risk. So while it is possible that brain chips could be widely adopted, it is important to bear in mind the possibility that implants may be limited to therapeutic uses and may not be widely adopted for cultural or pragmatic reasons. Further, if machine consciousness isn't technologically feasible, microchips may not function in parts of the brain responsible for consciousness without causing a loss of conscious activity (Schneider [2019b]).

It should be noted, however, brain organoid research could succeed where silicon or

[11] See https://neuralink.comhttps://neuralink.comhttps://neuralink.com.

alternate nonbiological materials do not, stepping in with AI trained biological brain implants. My earlier concern about whether microchips are the "right stuff" for consciousness would not apply to the case of "chips" that are implanted brain organoids. Such will not supply all the same sorts of augmentation, however, such as wireless internet connections or other capabilities nonprogrammable into biological entities.

Transhumanists such as Elon Musk, Nick Bostrom and Ray Kurzweil have urged that humans will not just use implantable chips, but will eventually upload their brains, merging with something like a superintelligent GB network and reaching a kind of technotopia that is depicted as an explosion of intelligence and heightened consciousness (Kurzweil [2005], Bostrom [2014], Schneider [2019a]). But it remains to be seen whether uploading will ever be scientifically feasible, and even if a brain's connections can somehow be copied by a program, this does not entail that one can 'transfer' their consciousness or identity to another substrate, so this technology may not be adopted over biological enhancements (see Schneider [2019a]). But we can put issues like uploading aside for the purpose of the GB argument, for you may already be a d-node in an incipient hyperintelligent system simply by using your digital devices and logging on to social media platforms.

Social media

Consider what social media platforms do to succeed financially, a phenomenon which has increasingly received widespread attention (Lanier [2018], Ward [2022], Frischmann and Selinger [2018]).[12] To maximize profits, platforms like TikTok and Facebook expose people to as much emotional information as possible, because doing so maximizes the amount of time a user spends on the platform in a single visit and encourages repeat visits. When a user is on a social media platform and consumes negative information, different networks of their brain compete, interact and cooperate. While the natural tendency is for individuals to utilize emotion-based learning processes when exposed to stressful or threatening information, these brain networks can be pitted against other networks in the brain that instantiate deliberation, self-regulation, and learning. Whichever network "wins" dictates how the information is then used for subsequent belief formation and behaviors. Social media algorithms that feature continuous, rapid-fire bits of emotionally evocative information prompt the activation of reward and emotion-based brain networks to facilitate non-conscious learning (*i.e.*, "associative" and "emotion-based

[12] This is due to the groundbreaking documentary (free on Netflix) called "The Social Dilemma", as well as the work of Center for Humane Technology and scholars such as those cited above.

learning"). Information learned through these channels tends to be more vivid and long lasting, and it has outsized influences on confirmation biases, tending to reinforce the beliefs that the individuals already have. The information learned in this way also bolsters availability heuristics (*i.e.*, one's assuming something happens at a much greater frequency than base-rates actually suggest) and aversion-based perceptions and behaviors towards others, encouraging fear-based perceptions of "us versus them" and fostering group polarization.

Processing in these networks often competes with that of other brain networks like the frontoparietal (FPN), default mode (DMN) and hippocampal-based networks that are involved in more conscious, self-directed learning. Such processing opens individuals to critical thinking and encourages more lasting attitude change. It is these networks that are essential to our ability to think rationally, evaluating whether the information we receive when engaging with a chatbot or other algorithm is reliable or truth conducive (in the case of externalist justification) and to provide and explain our reasons we used when arriving at a belief, in the case of internalist justification.[13]

While our data goes to the social media networks, the algorithms provide feedback to us, so we function as dynamic nodes. While loops need not lead to negative consequences, sadly, the kind of social media feedback loops humans engage in have had dire consequences for users. Sadly, today's social media platforms engage brain dynamics in a way that facilitates confirmation bias, with fear and emotion being primary engines for nurturing it. Users of these platforms are routinely encouraged to (if not only provided with a means to) privilege their existing beliefs, and are rarely, if ever, presented with evidence that disputes those beliefs (Ward [2022]).

When social media presents individuals with a continuous stream of information that evokes negative emotional responses, these learning contexts provide the ingredients for the well-documented effect in the cognitive neuroscience literature of emotional memory encoding (Orlowski [2020]). A large body of research on this topic illustrates that negative, emotionally charged information receives privileged attention and is better encoded, consolidated and retrieved in negatively arousing and stressful contexts (*e.g.*, Hamann [2001], LaBar and Phelps [1998], Ochsner [2000] Payne *et al.* [2006], Levine and Burgess [1997]; for a review see LaBar and Cabeza [2006]). Further, the effects are more enduring and vivid (Canli *et al.* [2000], Hamann *et al.* [1999]).

The upshots of this body of research on social media are important for the GB argument.

---

[13] I am grateful to Steven Gubka, Garrett Mindt and Chad Forbes for discussion of these issues.

Even if chips are not adopted as widespread brain enhancements, there is an important sense in which humans are already functioning as D-nodes in today's AI services and networks connecting them (Schneider [forthcoming]). Below, I further comment on the social implications of this. While organizations are explaining this phenomenon to the general public, and parents are becoming increasingly aware of the impact this has on the young, a new wave of dangers has emerged. For AI chatbots are being integrated into these platforms as well. Should chatbots draw from these techniques, this could deepen the manner in which individuals are manipulated and networked into the system.

I have now explained the substantive premises of the GB argument. But what does this mean for the future of humanity? Will the global brain be good or bad for human flourishing, perhaps leaving us in an Orwellian situation? Is a GB here, and if so, to what degree? (I have not fully defended Premise Two, because I believe it is difficult to tell if it is correct. More on this shortly.) What are the safety risks? Is this a kind of extended cognitive system, and do those d-nodes who are "wired in" have extended minds? These matters are the task of the remainder of this target article.

5 Is the global brain good or bad?

There's an older global brain hypothesis, and it tends to be utopian, for like Wells and Orwell, the proponents simply didn't anticipate the ills of AI alignment, cybersecurity threats and social media bubbles (Heylighten [1997], Geortzel [1998]).

As Ben Goertzel explains:

> How much more exciting, then, is the birth of an *entirely new kind of intelligent organism.* And this is precisely, I suggest, what those of us who are alive and young today are going to have the opportunity to experience: the birth of an entirely new kind of intelligent organism, a *global Web mind.* The Internet as it exists today is merely the larval form of something immensely more original and exciting. [Geortzel n.d.]

Proponents of the older global brain view tends to say that the internet will increasingly link its users into a single information processing system that functions like a collective self-aware system, a nervous system at a planetary level. However, I'm not committed to the idea that the entire internet will be something like a hyperintelligence. Nor am I claiming that individuals will upload their minds or that a global brain would have feelings, consciousness or self-awareness, indeed, elsewhere I've argued for a "wait and see" approach

to AI consciousness and have been highly critical of uploading claims (Schneider [2019a]).

This being said, let us now ask: will the global brain be good or bad for us? It is difficult to predict how such a complex phenomenon will play out, for AI development depends upon scientific developments, geopolitical considerations, regulatory behaviors and AI alignment efforts. But bearing this in mind, a serious concern is that today's growing AI chatbot ecosystem presents a "slow boiling frog" problem (Schneider [forthcoming]). According to the metaphor, (and it is just a metaphor), if you boil a frog by putting it in scalding hot water, it will try to save itself. If you put the frog in a pot of tepid water, it will not notice it is boiling so it will not try to save itself.[14] In both cases, the outcome is the same—the frog dies. In a similar fashion, I am concerned that a combination of factors, over time, gives rise to unhealthy engagement with chatbots and ultimately, to diminished human agency (Schneider [forthcoming]). The factors include:

- Epistemic deficits in LLMs (opacity, hallucinations, etc.).
- Impoverished digital privacy.
- Relationships with personalized chatbots, which people see as "digital companions" or "digital persons", blurring the lines.
- Epistemic trust in these 'digital companions' despite their epistemic deficits.
- Social media companies using principles in social psychology and neuroscience to manipulate chatbot users.
- The AI Megasystem Control problem.

First, we've seen that the chatbots have epistemic problems such as hallucinations, opacity and bias. Second, increasingly, people seem to be using chatbots as advisors and as relationship companions, perhaps thinking the bots have feelings and may be sentient, and despite the documented lack of data privacy. Third, as sketched in the present section, even before chatbots were integrated into social media platforms, humans were being manipulated by social media networks. While parents and educators are becoming increasingly aware of

[14] The slow boiling frog case is just a cultural metaphor for group or individual inaction due to people gradually getting used to a phenomenon that slowly increases, apparently a frog that is heated gradually will jump out.

the impact the use of platforms like TikTok has, people nevertheless persist in using the platforms, and many users are still ignorant of these issues or they simply do not care.

Now consider interacting with a chatbot that utilizes these same techniques the social media companies have already used to persuade users of 'truths' and keep users engaged. First, consider a chatbot on a social media platform presenting itself as a human user, what Dennett called a "counterfeit human" (Dennett [2024]). This could be on a platform in which the AI chatbot creators somehow have access to the personalized details of an individual, increasing the likelihood of manipulation. Second, consider a situation that involves a personalized AIs that a user is intentionally using regularly for information, advice and perhaps a more intimate connection. In this case, the AI is able to use the above techniques to manipulate the belief system of the user, but the user knows it is engaging with a bot. In both cases, the bots are able to be more persuasive by employing the well-known tactics sketched in the previous section.

When the water boils, the frog dies, or so goes the metaphor. In the human case, how might all these phenomena impact human flourishing? The regulatory, economic and political ecosystem that will serve as the backdrop for the AI developments over the next several years is, of course, still unfolding, and, again, there will inevitably be "unknown unknowns" along the way. But bearing this in mind, my own assessment is that if little or nothing is done to improve this situation, human agency can be compromised by the combination of factors I've identified in at least the following ways.

Humans will increasingly rely on chatbots for intellectual work and personal advice and connection, despite the aforementioned drawbacks with LLMs (hallucinations, opacity, etc.) which are well-known by the AI community. Relatedly, others worry that as AIs take over tasks that humans normally do that involve creativity and analysis, humans may less frequently use and develop creative, analytical abilities. (Slattery *et al.* [forthcoming]) Further, humans may increasingly avoid human relationships for relationships with chatbots. All the while, principles of human brain function can be mined by producers of chatbots to manipulate users, as has happened in the context of social media. Personalized chatbots that know exactly how to persuade someone, and a person's needs, are all the more powerful. And all the while, AI companies can mine the user interactions for data. The manipulations can in principle include injecting political or other kinds of bias.[15] So perhaps Orwell was

[15] As time progresses, the younger generations will inevitably not personally remember a time before society was tethered to social media and smartphones. Many have digital lives in which they hand over their personal details with little or no concern. Today's children may grow up with chatbots that are a close part

more on target than Wells, after all (Schneider [forthcoming]).

This said, the concept of a global brain is not intrinsically bad. For example, perhaps, as time goes on, a GB turns out to be more akin to Wells' democratic dynamical global encyclopedia, and the system it is based upon is one which respects user privacy, does not infringe on copyright protection, and avoids other dystopian scenarios. And perhaps this system also serves as a global sensor system that protects humans from natural disasters, biosecurity threats, and informs climate change research. Yet these same capacities are also compatible with a control structure in which a surveillance state has the perfect panopticon, both knowing your every move and making evident to you that it knows your every move. This knowledge that you are surveilled is enough to keep you in line (Foucault [2009]).

A GB is a *dual use technology*—a technology which can be developed for both civilian and military (or more broadly, coercive) purposes. For instance, one or more AI megasystems could seed misinformation during a wartime conflict, empower an overstepping regime, or gather data on a dissident population. And even if a dominant AI corporation aims for its proprietary megasystem to be apolitical, for business reasons, there are still ideological views that enter in during the process of RLHF—corporate interests involving data collection and discouraging competition, the slanting of discussions for what they believe to be security, political correctness (or deliberate efforts to avoid it) and/or safety reasons (See Schneider [forthcoming]). Just as today's social media platforms recruit brain networks to addict users to their platforms, so too, a GB may manipulate user psychology in highly individualized ways, such as through tailored personalized chatbots that anticipate and exploit our psychological weaknesses. And worse yet, autonomous weapons, connected to world state sensor systems and faulty and/or misaligned AIs could trigger a global disaster (Bailey [2025]).

There may be, for all we know, a global brain in existence today. While it is too early to tell to what extent humans will integrate with the global brain, it is clear that many humans are already in feedback loops with social media algorithms, and that the AI ecosystem is increasingly populated with generative AIs and more and more sophisticated AI megasystems. In sum, there is a very real possibility that we may be forced to participate in hyperintelligence, but rather than realizing a Kurzweilian or Wellsian technotopia, we have a decreased quality of life.

One may think these are the ingredients for the perfect panopticon of a techno-

---

of their lives, helping them with their homework and relationship problems.

authoritarian state, achieving a route to centralized knowledge through its tracking of the activity of nodes, but an authoritarian state will want tight control over a GB network, keeping it leashed in (type 5ii). But a GB system may, over time, become agential (type 5i). As I explain in the subsequent section, megasystems can quickly spin out of control. In sum, the future of the global brain depends upon us. Over time, our domestic regulations and international relations on the AI front may foster human flourishing or fail humans entirely, leaving us all in a common place—perhaps in an inadvertent AI generated nuclear disaster, an AI driven pandemic, or even like the *Borg* on *Star Trek*.

6 A new control problem—the problem of AI megasystem control

Thus far, I have been concerned with the intelligence of AIs, such as LLMs in their current iterations, as single systems, albeit systems that may indeed exhibit improving levels of intelligence, and perhaps further phase transitions and erratic behaviors, as the tech companies develop the models further. But there is another sense in which AIs will evolve. While the world's attention remains at the level of single AI systems like GPT-4, it is important to see where all this may be headed. Given that there is already evidence of erratic and autonomous behaviors at the level of single AI systems, what will happen when in the near future the internet becomes a playground for thousands of increasingly intelligent LLM systems, widely integrated into search engines and apps, interacting with each other? (See Schneider and Kilian [2023])

The classic control problem of AI concerns a scenario in which a system outthinks humans, becoming "superintelligent", and, because it is superintelligent and can outthink us, we lose control over it (Bostrom [2014]). What Schneider and Kilian [2023] call the "Megasystem Control Problem", concerns how to control the behavior of AI "megasystems". The megasystem control problem is the following problem: many elements of the AI ecosystem are designed by different organizations and the different systems may not align with each other, due their being developed by different and even competing organizations and their having unpredictable (weakly) emergent features of their interaction, including the efforts of human users to subvert parts of the internet ecosystem (Schneider and Kilian [2023]). Since unforeseen features already emerge at the scale of a single chatbot system, it would be shortsighted to ignore the possibility that erratic and autonomous behaviors could emerge from a megasystem comprised of interactions among a range of chatbot and other

generative AI systems at the level of the internet ecosystem.

I've already observed that research on multi-agent AI interaction illustrates that AIs can quickly evolve a secret language and manifest power-seeking behaviors. For example, during a simulated game of hide-and-seek, OpenAI observed two AI teams stockpiling objects from the environment to gain advantage over the competing team, what OpenAI described as being a form of "emergent tool use" (OpenAI [n.d.]). While one reaction to the hide and seek example may be to breathe a sigh of relief because the AIs only compete against each other, not humans, this misses the point. For the game was just a circumscribed environment involving just AIs, and real world AI systems will have concrete human impacts (Kilian et al. [2023], Schneider and Kilian [2023], Bailey [2025]).

Moving beyond a simple game, consider that today's Internet is increasingly seeing intelligent chatbots widely integrated into search engines and apps. They are currently being developed to be in intense competition with each other, by actors like Microsoft, Alphabet, or the US and Russia or China. This is a "Wild West" of vastly more competitive and complicated interactions than a simple game of hide and seek. With such rapid-fire advances in the chatbot ecosystem, it is crucial to anticipate how the AI services can themselves self-organize into alignments, warring factions and potentially coalesce into emergent ultra-intelligent systems with behaviors harmful to humans, forming increasingly sophisticated AI megasystems (Kilian *et al.* [2023], Schneider and Kilian [2023]). It is not hard to imagine the wide-ranging consequences of a megasystem that hacks into critical systems, provides destabilizing information to the public, through chatbots or social media, or produces and distributes instructions for the next megavirus.

So, what safeguards can be put in place? Companies such as Microsoft, Anthropic and OpenAI are diligently developing means to deal with AI behaviors at the level of their particular products. But as chatbots like GPT-4 increase in scope and size, they evolve new features that were not present in earlier versions of the model. These have been called "emergent" abilities in the AI literature, a capability or behavior that arises spontaneously or unexpectedly from the interactions or complexity of a system's components that was not present, or at least not detected, in earlier versions of the program. Although this may merely be an epistemic phenomenon, arising due to our inability to adequately gauge the system's capacities at a given time, it is still worrisome. For in either case, it indicates human inability to understand how a system will behave when on the larger internet.[16]

---

[16] We shouldn't assume that all cases of LLM emergence are due to the same underlying phenomenon. There is

(Wei *et al.* [2022], Schneider and Kilian [2023], Bailey and Schneider [2023]).

A concern for unforeseen behaviors is why we see companies putting out a limited or a cautious supervised release of their AI chatbots. The companies then react to the feedback by altering certain characteristics, such as the behavior of ChatGPT's Sydney alter ego. A cautious release of a chatbot can put the brakes on an erratic model version like Sydney and it can help with the traditional Control Problem—the challenge of controlling a single AI system that could, in principle, outpace our ability to control it (Bostrom [2014]). However, in the context of the behavior of one or more megasystems on the AI ecosystem, this is the equivalent of focusing on a single bird to explain flocking behavior. The range of AI services on the Internet is not owned by a single organization, of course, and different AI services are designed to compete. So no single corporation or government has the capacity or authority to control the behavior of a hybrid AI megasystem. To compound the situation, much more data and computational power are encompassed at the megasystem level, creating complex conditions for unforeseen interactions.

Where might all this lead? In the context of today's discussion the following concern arises: In addition to using social media platforms that recruit brain networks to foster addiction to their platforms, we may unwittingly embark upon a future where human beings, through their devices, are "wired" into a GB network consisting of proprietary AI services that include social media platforms and other services linked to a particular tech company, such as the constellation of AI services owned by Google. To be seamlessly integrated with AI services may not only have a negative impact on oneself (Schneider [2019a], Turner [2022]), but it may lead to AI Megasystems that are beyond the control of any one person or group to control.

We've seen that today's deep learning algorithms are often difficult for even the programmers themselves to understand. And while the black box problem is well-known, I foresee another problem, which I call the 'network identification problem'. The problem is that with emerging megasystems it will be difficult for ordinary users, and indeed, specialists, to ascertain where an information processing network that one is 'wired into' begins and ends. For consider that a single app can be used by multiple AI networks, each of which does different things with that data on their network. As such, the app is a subroutine or node in that larger network. Different AI networks will overlap in their parts. Further, if

---

also an important debate over whether emergence is merely epistemic, merely seeming the system undergoes phase transitions due to our inability to correctly measure the LLM capacities and behaviors. In either case, this is a deep challenge for predicting future capacities and behaviors of the models.

organizations or hidden actors or AI agents/factions are part of a network, a key part of the network may be unknown to the researchers. As a result, if users become engaged with chatbots on the megasystem their own cognitive, emotional and perceptual lives could be shaped by unidentifiable, emergent intelligences or malicious actors on the megasystem.

7 Is a global brain here?

An incipient global brain may be here already, being beneath our radar. AI networks are as unlike biological intelligence as they come. Unlike brains and nervous systems, the internet is decentralized and is not located in a continuous part of spacetime or inhabited by a single body, being spatiotemporally scattered by fiberoptic cables and across continents. We are creatures with brains and nervous systems and from our inner experience and interactions with others, we have an emotional and conceptual understanding of human physical and mental capacities, but the abilities of the AI systems are in many ways alien to us, despite the fact that we built them.

Relatedly, the scale is different. Consider the perspective of a single ant in a larger colony, lacking a conception of the workings of the colony. Both the ant and the colony are forms of intelligence in their own right. Of course, humans are capable of theorizing about complex phenomena that consist in the activities of groups of individuals, consider theories of markets, ecosystems and democracies, for example. And while no single human has a full understanding of macroeconomic theory, groups of humans know a good deal, developing theories, abstractions, methodologies. But to do this there needs to first be an appreciation that the phenomenon is a form of collective behavior in its own right, and what the challenges are. Here, the terrain is new.

For one thing, we lack a theory of evolution that applies to this form of intelligence. Unlike biological life, global brain networks do not evolve according to Darwinian evolution. We are hunter gatherers who have an evolutionary lineage beginning about 3.7 billion years ago from life's origin, to the achievement of highly complex beings with brains and nervous systems. Along the way, we achieved consciousness and the capacity to have a sense of self and to feel a range of emotions. and we became social beings. The global brain network evolves too, but its evolution is determined by factors like market forces, constraints on computational speed and cost, internet regulations and more. This is not a Darwinian process, and a GB may never hit the milestones we did, and it may invent new ones.

It's hard to tell for a further reason as well. I've noted that both LLMs and a GB network are dual use technologies. AI developments are important to the security of numerous

nations due to the import of research in quantum computing, cybersecurity and weapons systems. While we can observe cognitive capacities of AI services like Microsoft Bing, we do not know about the capacities of what I call "shadow megasystems", systems that draw from the contents of public megasystems, global state sensor systems, their proprietary supercomputers, gathering and synthesizing mounds of data for security purposes. For instance, a government may have access to the data that AI and social media companies utilize and produce. It would be naive to believe, given the import of information gathering to the security of various countries, that this information would not be of interest. Efforts of various countries would make use of existing megasystems, synthedized and analyzed by their own AIs, making these networks in effect part of shadow megasystems.

How causally integrated are these systems? Do some of them contain information drawn from shadow megasystems they hack into? How intelligent are they? Since these are shadow megasystems, I cannot answer these questions, nor can I answer the question: is there a global brain network today? If there *is* a shadow megasystem, let's hope it is "leashed in", although agential subsystems that can carry out certain tasks autonomously is also concerning, and even a leashed in GB system, in the wrong hands, is also dangerous. Further, an agential system could emerge, due to factors such as phase transitions and misaligned subsystems. Shadow megasystems are a risky business indeed. In sum, while we may not definitively know whether a global brain exists today, the convergence of advanced AI systems, integrated networks and human participation suggests that we are witnessing the early stages of its formation. So while it may frustrate the reader to see the advocate of an argument unwilling to firmly say that she believes her argument is true, I am at least ready to commit to the following: if premises two is not yet true, barring unforeseen circumstances, it will be correct within the next five to ten years.

8 The extended mind hypothesis and the global brain

Now let us turn to the question: Is a human that is a d-node a case of an extended mind? The Extended Mind (EM) hypothesis is the proposal that the physical substrate for the mind is not restricted to the central nervous system, but can include external physical items, as when one's memories are stored in external media such as notebooks and hard drives (Clark and Chalmers [1998], Clark [2008]).

Although "mind" is a term of art, many would say that having the capacity for conscious states is a necessary condition on something's having a mind or being minded. This may lead one to assume that proponents of EM believe that consciousness is extended. But many

advocates of the EM hypothesis are not in fact saying that consciousness is extended. So herein, I call the view the "extended cognition" (EC) hypothesis, rather than the "extended mind".[17]

Now, when a human functions as a node in a GB network, is their cognitive system extended? Certain critics of EC have objected that laptops and notepads do not exhibit a sufficiently rich cognitive integration with the brain to justify EC. For one thing, information from notebooks and laptops enters into cognitive systems through inputs to sensory transducers and can impact the brain's cognitive processing, but critics claim that the information is not sufficiently integrated into the brain, functioning like the information flow involved in actual brain processes themselves (Adams and Aizawa [2001, 2008]). Further, critics can claim that if an unimpaired human forgets their laptop or notebook, the brain is still able to function, without confusion, impairment or disability when it lacks access to these devices, suggesting that the external components were not really part of one's cognitive system to begin with. Of course, Clark and Chalmers have provided a well-known example of a person with Alzheimer's disease (named "Otto"), who needs a notebook to function, and even unimpaired humans can be quite confused without access to their GPS.

I will not try to settle these debates herein. But if these critics are correct, a human that is a d-node *isn't* a case of EC. A human d-node can be in dynamical interaction with the global brain, even being manipulated by it, but the integration is not extensive or rich enough to constitute a case of EC. So even if the critics of EC are right, humans can still be a d-node in a GB network if they are not cases of EC. My own feeling is that simply being a d-node through using a laptop is not a case of EC, but other d-node instantiations could, in principle, have greater integration and even perhaps satisfy these critics. In particular, Joseph Corabi and myself have argued that brain implants could, with technological development, become sufficiently well-integrated with the biological brain to be cases of EC, for the implants could in principle function as part of minicolumns or brain regions. Indeed, back in 2008 Andy Clark, following up on his 2003 book, had raised a hypothetical neural prosthetic case in a published letter that was his response to Jerry Fodor's critical review of his book, *Supersizing the Mind* (Fodor [2009]). The case involves a woman, named "Diva", who has brain damage and can no longer perform division. A silicon microchip is added to her brain, restoring her previous ability. As she computes,

[17] We might also include perception here. But in this brief discussion I will not include it.

a mental process is implemented by a hybrid biological and silicon-based system. So, is Diva's mind extended? Clark concluded:

> That alone, if you accept it, establishes the key principle of *Supersizing the Mind.* It is that non-biological resources, if hooked appropriately into processes running in the human brain, can form parts of larger circuits that count as genuinely cognitive in their own right. [Clark 2009]

This seems correct. If brain implants become more widespread, neural prosthetics could be a common and far less controversial way in which we are cyborgs. For, to use the case of Ted Berger's artificial hippocampus as an illustrative example, given that the hippocampus is part of one's cognitive system, it seems implausible to deny that an artificial hippocampus, which is a device playing a similar (if more basic) functional role as the hippocampus, isn't part of one's cognitive system as well. Here we could see a kind of tight cognitive integration and speed of information flow that critics found lacking in the notebook and smartphone cases, and resembles the way the brain interacts with its own parts (Schneider and Corabi [2021]).

You might object that this case doesn't extend the mind outside of the body. However, the artificial hippocampus project, to the best of my knowledge, *is* currently a project in which the electrodes link to an actual processor that *is* outside of the body. (However, the long-term goal of the research program is to locate the device entirely *inside the head*.) Further, for any sort of neural prosthetic, it seems *prima facie* plausible that if science can create an artificial part that is located in the head, then, as Clark [2008] had observed in the Diva case, the device could be outside of the head and even the body, communicating wirelessly. At some point in the future, a remotely connected device has the potential to bear similar richly networked connections to the brain (Schneider and Corabi [2021]).[18] This suggests the case for EC is strong with integrated implants. So a d-node that has an integrated chip in a GB network seems to be a relatively clear case of EC, whereas the earlier case of digital devices and laptops is more controversial. The GB argument does not entail EC, then, at least without an argument that d-nodes without implants can be cases of EC.

Intriguingly, in either case, the following situation seems true of the human that is a

[18] Work on neural prosthetics is important to the EC hypothesis for a second reason: the extended cognition hypothesis can be turned into something of a testable hypothesis--a hypothesis which can be confirmed by the successful use of neural prosthetics to underlie cognitive functions in the brain, or outside of the brain. As such, it is a means of making tangible progress in the debate over EC (Schneider and Corabi [2021]).

d-node:

> **System externalism**: a human (a d-node) may not know the nature of the larger distributed cognitive system of which he or she is a part.

Earlier, I discussed a 'system demarcation problem' with respect to the GB network. Because a single megasystem can encompass a range of AI services owned by different organizations and likely includes non-public systems it is difficult to know where a system begins or ends, even for a team of AI experts. Now consider how this situation impacts the epistemic situation of a d-node: not only does one not know the full nature of the extended system, but one doesn't even know *how many* global brain systems one is a d-node in. (Since one or more GB networks can provide feedback and control of one's cognitive system, *prima facie*, this may strengthen the case for content externalism, as a superior explanation for a d-node's behavior, but I will not pursue this issue today.)

Further, if one is a d-node, the process of epistemic justification may be tainted. First, we've seen that today's LLMs are not reliable systems, for they hallucinate, have a capacity to be biased, are sycophantic and more. It is also not clear that LLMs will ever be free of hallucinations, and it may be that sycophantic chatbots persist because they get more users. In addition, suppose instead an internalist framework for epistemic justification in which justification involves internal reflection on the operative beliefs. Because the individual that is a d-node cannot determine the extent or nature of the larger system for the information, it is difficult to see how they can assess its epistemic strengths and weaknesses, indeed, this information may not even be available by asking experts (Schneider [forthcoming]).

9 Conclusion

In this paper I've laid out the Global Brain argument and explored some of its implications, asking: what does this mean for the human future and the future of human-machine interaction? Is a global brain here? What are the AI safety implications? Are those who are "wired in" cases of extended minds? While a single paper cannot explore all of the implications of the argument, or discuss them in the depth they demand, I hope this article inspires further reflection. I believe that the future of human flourishing demands that we consider the implications of this emerging intelligence. We must study the phenomenon with diligence and foresight, shaping its capacities and managing its impact, being mindful that this intelligence may shape the future of life on Earth.

Susan Schneider
Florida Atlantic University
Center for the Future of Mind, AI and Society, SO 278
777 Glades Road
Boca Raton, FL 33431-0991
susansdr@gmail.com

References

Adams, Frederick & Aizawa, Kenneth [2001]. "The bounds of cognition". *Philosophical Psychology* **14**: 43–64.

Adams, Frederick & Aizawa, Kenneth [2008]. "The bounds of cognition: A reply to Clark". *Philosophical Psychology* **21**: 589–602.

Anthropic [2023]. *Claude 2*. https://www.anthropic.com/index/claude-2.

Bailey, Mark [2025] *Unknowable Minds*. New York: Imprint Academic.

Bailey, Mark & Schneider, Susan [2023]. "AI shouldn't decide what's true". *Nautilus* (2023, May 18). https://nautil.us/ai-shouldnt-decide-whats-true-304534/?_sp=c90c9253-da9f-4e7f-b5cc-da8f5347f9e0.1684413883717.

Bostrom, Nick [2014]. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bubeck, Sébastien, Chandrasekaran, Varun, Eldan, Ronen, Gehrke, Johannes, Horvitz, Eric, Kamar, Ece, Zhang, Yi, *et al*. [2023]. "Sparks of artificial general intelligence: Early experiments with GPT-4". *arXiv preprint arXiv:2303.12712*.

Canli, Thuran, Zhao, Zuo, Brewer, James, Gabrieli, John D., & Cahill, Larry [2000]. "Event-related activation in the human amygdala associates with later memory for individual emotional experience". *Journal of neuroscience 20*: RC99.

Chalmers, David [2006]. "Strong and weak emergence". In *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, edited by P. Clayton & P. Davies. New York: Oxford University Press: 244–54.

Clark, Andy [2008]. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Clark Andy [2009] "Letter on Fodor on "Where Is My Mind?"". *London Review of Books*.

Clark, Andy & Chalmers, David [1998]. "The extended mind". *Analysis* **58**: 7–19.

Dennett, Daniel C. [2023]. "The problem with counterfeit people". *The Atlantic*, May 31. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-

people/674075/

Farquhar, Sebastian, Kossen, Jannik, Kuhn, Lorenz & Gal, Yarin [2024]. "Detecting hallucinations in large language models using semantic entropy". *Nature* **630**: 625–30.

Fodor, Jerry A. [2000]. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, Mass.: MIT Press.

Fodor, Jerry A. [2009]. "Where is my mind?". *London Review of Books* **31**: 13–5.

Fodor, Jerry & Pylyshyn, Z. W. [1988]. Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Foucault, Michel [1995]. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan. New York: Vintage Books.

Frischmann, Brett, & Selinger, Evan [2018]. *Re-engineering Humanity*. Cambridge: Cambridge University Press.

Germain [2023]. "When AI gets it wrong: Addressing AI hallucinations and bias". https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

Goertzel, Ben [1998]. "World wide brain: Self-organizing Internet intelligence as the actualization of the collective unconscious". In *Psychology and the Internet: Intrapersonal, interpersonal and transpersonal implications*, edited by J. Gacken-bach. New York: Academic Press.

Goertzel, Ben [n.d.]. *World Wide Brain: The Emergence of Global Web Intelligence and How it Will Transform the Human Race*. https://www.goertzel.org/papers/webart.html.

Goertzel, Ben & Pennachin, Cassio (eds.) [2007]. *Artificial General Intelligence*. Cham: Springer.

Google Deep Mind Gemini Team [2023]. "Gemini: A family of highly capable multimodal models". Preprint at https://arxiv.org/abs/2312.11805.

Hamann, Stephan [2001]. "Cognitive and neural mechanisms of emotional memory". *Trends in Cognitive Sciences* **5**: 394–400.

Hamann, Stephn B., Ely, Timothy D., Grafton, Scott T., & Kilts, Clinton D. [1999]. "Amygdala activity related to enhanced memory for pleasant and aversive stimuli". *Nature neuroscience* **2**: 289-93.

Heylighen, Francis [1997]. "Towards a global brain. Integrating individuals into the world-wide electronic network". In *Der Sinn der Sinne*, edited by U. Brandes & C. Neumann. Gottingen: Steidl Verlag.

Holland, John H. [1995]. *Hidden Order: How Adaptation Builds Complexity*. Reading: Addison-Wesley.

LaBar, Kevin S. & Cabeza, Roberto [2006]. "Cognitive neuroscience of emotional memory".

Nature Reviews Neuroscience **7**: 54–64.

LaBar, Kevin S. & Phelps, Elizabeth A. [1998]. "Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans". *Psychological Science* **9**: 490–3.

Lanier, Jaron [2018]. *Ten Arguments for Deleting Your Social Media Accounts Right Now*. New York: Random House.

Levine, Linda J. & Burgess, Stewart L. [1997]. "Beyond general arousal: Effects of specific emotions on memory". *Social Cognition* **15**: 157–81.

Llama Team [2024]. "The llama 3 herd of models". https://arxiv.org/abs/2407.21783.

Kahneman, Daniel [2011]. *Thinking, Fast and Slow*. New York: Macmillan.

Kurzweil, Ray [2005]. *The Singularity is Near: When Humans Transcend Biology*. New York: Viking Press.

Kilian, Kyle A., Ventura, Christopher J & Bailey, Mark M. [2023]. "Examining the differential risk from high-level artificial intelligence and the question of control". *Futures* **151**.

McCulloch, Warren S., Pitts, Walter [1943]. "A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics* **5**: 115–33.

McLaughlin, Brian [1992]. "The rise and fall of British emergentism". In *Emergence or Reduction?: Prospects for Nonreductive Physicalism*, edited by A. Beckermann, H. Flohr & J. Kim. New York: De Gruyter: 49–93.

Mok, Aaron [2023]. "What do AI art generators think a CEO looks like? Most of the time a white guy". Business Insider. Accessed March 28. https://www.businessinsider.com/ai-art-generators-dalle-stable-diffusion-racial-gender-biasceo-2023-3.

Mitchell, Melanie [2009]. *Complexity: A Guided Tour*. Oxford: Oxford University Press.

Nguyen, C. Thi. [2020]. "Echo chambers and epistemic bubbles". *Episteme* **17**: 141–61.

Nicoletti, Leonardo & Bass, Dina [2023]. "Humans are biased. Generative AI is even worse". Bloomberg Graphics. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

Ochsner, Kevin N. [2000]. "Are affective events richly recollected or simply familiar? the experience and process of recognizing feelings past". *Journal of Experimental Psychology: General* **129**: 242–61.

OpenAI [n.d.]. "Emergent tool use from multi-agent interaction". https://openai.com/research/emergent-tool-use.

Orlowski, J. (Director) [2020]. *The Social Dilemma* [Film]. *Exposure Labs*.

Orwell, George [1950]. *1984*. London: Penguin Books.

Payne, Jessica D., Jackson, Eric D., Ryan, Lee, Hoscheidt, Siobhan, Jacobs, Jake, & Nadel,

Lynn [2006]. "The impact of stress on neutral and emotional aspects of episodic memory". *Memory* **14**: 1-16.

Porter, Tenelle, Elnakouri, Abdo, Meyers, Ethan A., Shibayama, Takuya, Jayawickreme, Eranda, & Grossmann, Igor [2022]. "Predictors and consequences of intellectual humility". *Nature Reviews Psychology* **1**: 524–36.

Roose, Kevin [2023]. "A conversation with Bing's chatbot left me deeply unsettled". *The New York Times*, February 16 2023.

Russell, Stuart & Norvig, Peter (eds.) [2020]. *Artificial Intelligence: A Modern Approach, 4th ed.*. New Jersey: Pearson.

Schneider, Susan [2019a]. *Artificial You: AI and the Future of Your Mind.* Princeton, NJ: Princeton University Press.

Schneider, Susan [2019b]. "Should you add a microchip to your brain? You might risk losing yourself". *New York Times*, June 10 2019.

Schneider, Susan [forthcoming]. "Chatbot epistemology". *Social Epistemology*.

Schneider, Susan & Corabi, Joseph [2021]. "Cyborg divas and hybrid minds". In *The Mind-Technology Problem. Studies in Brain and Mind, vol 18*, edited by Clowes, R.W., Gärtner, K., Hipólito, I. The Mind-Technology Problem. Cham: Springer: 145–59.

Schneider, Susan & Kilian, Kyle [2023]. "Artificial intelligence needs guardrails and global cooperation". *The Wall Street Journal*, April 28. https://www.wsj.com/articles/ai-needs-guardrails-and-global-cooperation-chatbot-megasystem-intelligence-f7be3a3c

Sharma, Mrinank, Tong, Meg, Korbak, Tomasz, Duvenaud, David, Askell, Amanda, Bowman, Samuel R., Cheng, Newton, Durmus, Esin, Hatfield-Dodds, Zac, Johnston, Scott R., *et al.* [2023]. "Towards understanding sycophancy in language models''. https://arxiv.org/abs/2310.13548.

Slattery, Peter, Saeri, Alexander, Grundy, Emily A. C., Graham, Jess, Noetel, Michael, Uruk, Risto, Dao, James, Pour, Soroush, Casper, Stephen & Thompson Neil [forthcoming]."The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence".

Testa, Bernard, Kier, Lemont B. [2000]. "Emergence and dissolvence in the self-organisation of complex systems". *Entropy* **2**: 1–25.

Turner, Cody [2022]. "Neuromedia, cognitive offloading, and intellectual perseverance". *Synthese* **200**: 66–92.

Wei, Jason, Tay, Yi, Bommasani, Rishi, Raffel, Colin, Zoph, Barret, Borgeaud, Sebastian, Yogatama, Dani, *et al.* [2022]. "Emergent Abilities of Large Language Models". Transactions on Machine Learning Research. https://openreview.net/

forum?id=yZrjA3vRxi.

Wells, H. G. [1938]. *World Brain*. London: Methuen.

Wiener, Norbert [1948]. *Cybernetics: Or Control and Communication in the Animal and the Machine*. John Wiley.