https://doi.org/10.1038/s41467-025-59906-9

Extending Minds with Generative AI

Andy Clark

Check for updates

As human-AI collaborations become the norm, we should remind ourselves that it is our basic nature to build hybrid thinking systems – ones that fluidly incorporate non-biological resources. Recognizing this invites us to change the way we think about both the threats and promises of the coming age.

Sometimes, it seems like the world is full of techno-gloom. There is a fear that new technologies are making us stupid. GPS apps are shrinking our hippocampus (or in other ways eroding our unaided navigational abilities); easy online search is making us think we know (unaided) more than we do; multitasking with streaming media is driving down our native attention span and (possibly) affecting grey matter density in anterior cingulate cortex - see refs. 1–4 and for a balanced review[5].

Similar fears surround new forms of generative AI (such as ChatGPT and others). These powerful tools – when properly prompted - can create novel versions of text, pictures, and other kinds of data, in ways that serve our needs. Will the ready availability of such tools encourage cheating and undermine educational practice[6]? Worse still, will it slowly rob us of both the need and the drive to create and value our own content? If we need to write a letter of complaint, plan a wedding, or compose an essay, perhaps we will simply fire up our favorite generative AI and uncritically accept whatever it offers? Wisdom suffers, collective self-plagiarism looms, and human creativity becomes all but obsolete.

When easy search, location services, constant access, and generative AI-based toolkits come together, they weave a complex web whose implications for humanity are far from clear. On the one hand, these webs empower us to do more and more, and more easily, than perhaps ever before. On the other, many fear they may rob us of more elusive goods. We might become content-curators rather than creators, and the passive targets of click-rate maximizing algorithms that do not have our best interests at heart. And what of human flourishing? Don't these developments threaten to rob us not just of our time and attention but the very joy of creation too?

Grim portents indeed. But there are other ways to picture these emerging landscapes. There is ample reason to be alert and cautious – we'll come back to all that shortly. But in thinking about the effects of all our new tools and technologies, we may often be starting from entirely the wrong place. The misguided starting point is an image of ourselves as (cognitively speaking) nothing but our own biological brains. An alternative vision, one that I have long championed, is that we humans are and always have been, what New York University philosopher David Chalmers and I call 'extended minds' – hybrid thinking systems defined (and constantly re-defined) across a rich mosaic of resources only some of which are housed in the biological brain[7].

Yet we humans seem strangely resistant to recognizing our own hybrid nature. Instead, we have long harbored a highly restrictive image of ourselves – one that leads very naturally to the worry that new tools and technologies will lead to mental diminishment. In Plato's *Phaedrus*, written around 370BC, we find a clear statement of the fear that new-fangled inventions such as reading and writing will have catastrophic effects on human memory. The fear was that these innovations would lead to lazy minds. We would start to think that we knew more than we really did, thanks to these cheap but superficial new means of non-biological storage and retrieval. Those fears seem laughable now. But we have not yet rid ourselves of their root cause – that deeply mistaken cognitive self-image. This is puzzling, especially since the history of our species is marked again and again by the transformative effects of newly discovered tools and practices. Arguably, it is our bedrock human nature to spread the load in this way – to become what I have called 'natural-born cyborgs'[8,9].

And yet, perhaps there was an element of truth even in the worries raised in the Phaedrus. Like most academics, I have often downloaded a paper and fooled myself into thinking I thereby – by some kind of electronic osmosis - knew what was in it. The advent of robust, reliably available, online search has added contemporary fuels to these familiar fires. Empirical studies have shown that the use of online search can lead people to judge that they know more 'in the biological brain' than they actually do, and can make people over-estimate how well they would perform under technologically unaided quiz conditions[2,10–12].

But looked at another way (as the authors of these studies mostly acknowledge) these mistakes can also seem like quite a small price to pay[5,10,11]. As our tools and technologies have progressed, we have been able to probe ever further and deeper into the mysteries of life and matter. We have come to understand much about the likely conditions at the very start of time and unlocked the biochemical foundations of life. We have not achieved this by becoming dumber and dumber brains but by becoming smarter and smarter hybrid thinking systems. Let's take a look at some basic elements.

The first stop, beyond the brain, is the rest of biological body. For example, we gesture as we communicate, and those physical gestures have been shown to reduce the cognitive load on the brain – a kind of spontaneous version of counting on your fingers[13]. Reaching just beyond the body, we exploit the rich world of familiar old-school static media. Even today, many of us will resort to pen and paper, scribbling madly as we try to think through a problem in math, life, or philosophy. More prosaically, we may sketch and re-sketch the design of a new kitchen on a paper napkin. These familiar loops through external media slowly become part and parcel of how many of us think. In my current work, I am part of a large interdisciplinary team (philosophers, archeologists, and vision scientists) exploring the many ways human-built worlds have acted, over the whole sweep of human history, as extensions to, and transformers of, human thinking. From stone tools to megaliths, and from marked trails to city plans, we humans have built worlds that extend our minds, and that alter the tasks that brain-based reasoning is called upon to perform.

To those old ploys we now add the burgeoning realms of VR, AR, AIs, apps for everything, and semi-intelligent devices. In all these ways (and endless more) we have been busily building ourselves new ways of

thinking and reasoning about our worlds and choices. As this process unfolds, the best uses of the biological brain (though not its fundamental operations and nature) once more shift and alter[3,5]. This is because human brains are amazingly adept at dovetailing their own native skills to the new opportunities made available by wave upon wave of tools and technologies. This results not in simple 'offloadings' of work so much as the creation of delicately interwoven new wholes - brain, body, world tapestries in which what the brain does, what the body does, and what the loops via external media and apps provide are all in continuous flux, each adapting (in its own way, and at its own timescales) to what the rest has to offer.

Recent neurocomputational work in the area known as 'predictive processing' (or 'active inference') has shown just how it is that brains like ours are poised to become woven into these larger tapestries of brain, body, and world. According to this emerging picture, we learn about our worlds by constantly trying to predict the sensory consequences of our own actions[14,15]. As we do so brains like ours become expert at resolving key uncertainties by taking different actions in the world. Such actions might include using a stick to probe the depth of a river – but they may also include firing up an online resource to help resolve some other kind of uncertainty[15,16]. When the world around us presents enriched suites of opportunities, brains like that learn to do the most efficient thing – for example, storing in bio-memory only the stuff (search cues, for example) that is needed to get the right results, as and when needed, from the larger ecologies in which they are situated.

The brain's core skill set thus includes launching actions that recruit all manner of environmental opportunity and support - from scribbling on a sticky note to firing up an AI. If the best uncertainty-minimizing suite of actions involves a bit of internal brainwork and a bit of bodily work (such as tapping some keys on a laptop) that is the sequence that gets chosen. The brain itself is unconcerned about where and how things get done. What it specializes in is learning how to use embodied action to make the most of our (now mostly human-built) worlds[17].

Still, there remains an intuitive difference between 'extending my own mind' through the use of pen, paper, and carefully-deployed apps, and (for example) simply asking someone else to solve the problem. In this regard, some of the most striking recent innovations – such as the use of ChatGPT and other generative AIs - might seem like rather bad candidates for extending human thought and reason. Instead of acting as mind-extending technologies, the fear is that these may act as mind-replacing technologies. How might we react to that kind of worry?

Here, the devil will (for some time at least) remain in the details. But there is suggestive evidence that what we are mostly seeing are alterations to the human-involving creative process rather than simple replacements. For example, a study of human Go players revealed increasing novelty in human-generated moves following the emergence of 'superhuman AI Go strategies'[18]. Importantly, that novelty did not consist merely in repeating the innovative moves discovered by the AIs. Instead, it seems as if the AI-moves helped human players see beyond centuries of received wisdom so as to begin to explore hitherto neglected (indeed, invisible) corners of Go playing space. The same will be true, I conjecture, in domains ranging from art and music to architecture and medical science. Instead of replacing human thought, the AIs will become part of the process of culturally evolving cognition. There, the relative alienness of the AIs thinking will sometimes work in our collective favor, enabling us to see beyond some of the prejudices and blind spots that have been hiding important new ways of thinking. However, the opposite effect can also occur, as noted (in the case of some bodies of scientific research) in another recent study[19]. That study alerts us to the potential role of AI in cementing certain tools, views, and methodologies in place, thus impeding the emergence of alternative approaches – much as, to borrow the authors' metaphor, an agricultural monoculture improves efficiency while making the crop more vulnerable to pests and diseases.

The lesson is that it is the detailed shape of each specific human-AI coalition or interaction that matters. The social and technological factors that determine better or worse outcomes in this regard are not yet fully understood, and should be a major focus of new work in the field of human-AI interaction. Encouragingly, the impact of AIs on the cultural variation and transmission of ideas is now being studied as a topic in its own right, already revealing a mixed pattern of good and damaging effects[20]. Such studies should help suggest targeted means of mitigating the bad and curating the good, by building infrastructure and legislation designed to offset our weaknesses and (as the authors neatly put it) keep human needs and human society firmly in the loop.

As part of this infrastructure, personalized AI-based resources are now emerging at the crucial interfaces between human users and the most powerful shared AIs. For example, there now exists 'Digital Andy'. Built by the computer scientist Paul Smart, Digital Andy applies a technique called Retrieval-Augmented Generation to the large language model (LLM) that drives a version of ChatGPT. What this means is that the ChatGPT base model has been augmented with an additional database of my own recent work. When a user queries Digital Andy, that query is then automatically augmented with relevant materials from the database so as to automatically enrich the prompts sent to ChatGPT. The responses generated are thus sensitive to changes in what I am currently thinking and writing, even when no texts reflecting those specific ideas were part of the training-data for the base LLM.

Even without such augmentation, the base model was capable of delivering plausible-sounding opinions on subjects that I have never directly addressed, such as the relations between 'extended minds' and 'quantum computing'. I think it quite possible that, as we get more used to working and acting within these kinds of (increasingly personalized) 'cognitive ecosystems', we will start to treat such suggested opinions in rather the way we might treat a thought that suddenly occurs to us during a conversation on some new topic. We treat the new thought as in some broad sense belonging to me[8]. But – just as with that sudden thought- we would also want to explore whether it really makes sense, and whether or not we are, all things considered, happy to endorse it. In this way, we both trust and question the suggestions of the more personalized resource, just as we might trust and question ideas that suddenly bubble up from our own biological unconscious.

Learning how to both trust and question our best AI-based resources in this way is one of the most important skills that our evolving educational systems will now need to install. New tools that help in assessing the outputs of standard LLMs will also play a role in that process. A nice example arose when researchers at Google DeepMind used an LLM to solve a famous long-standing (and important) mathematical puzzle[21]. The solution was entirely novel. But it did not emerge fully-formed by querying the LLM. Instead, it was gradually ushered into the world courtesy of a new tool called FunSearch (short for Function Space Search) that is expert at rejecting useless suggestions (the LLM made many), spotting the occasional, more promising

suggestions, and using those to repeatedly re-prompt the LLM until the solution was eventually discovered. New user-side tools like these may one day work together with other personalized interfaces to better position the generative AIs as true creativity-enhancing resources – welcome additions to how we think and decide, rather than simple, uncritically-deployed, replacements.

Another avenue for creatively synergizing with a persisting backdrop of apps, aids, and generative AIs looks set to involve the slow training of 'personal AIs'. Over time, these personalized resources learn about your own specific needs and interests. It is already possible to train a personal AI to share the workload in a specific set of tasks. One example is the GENIUS system from VERSES. Billed as a next-generation personal AI, the system learns (in multiple ways and across multiple platforms) from the user and offers highly-tailored suggestions and recommendations. In the future, these personally tailored AIs may form a kind of 24/7 ecosystem around you, the base biological agent. They may be trained by interacting with bio-you from a young age. They could learn from your projects and choices, and play a role in sculpting your future projects and choices. I think they would rapidly feel like 'borderline-you' - robust, reliably available, constantly running in the background, and implicitly trusted.

Whatever the technologies, such resources will swirl around our bio-cores, popping up ideas and opportunities, helping you do things, recognize situations, and achieve your practical goals. Some resources within the swirl will be able to sense stuff that bio-you can't sense. Some may be woven into intelligent clothing, which may in turn be monitoring your bio-states (of stress, anxiety, excitement etc.). You survive their loss or deletion, but much as you would a minor stroke. You may still count as 'using' these intimate digital technologies, but only in the same attenuated sense as you 'use' your hippocampus or frontal lobes[8]. Really, they are just part and parcel of the new hybrid you.

So what should we make of all that techno-gloom? Let's assume that the facts behind at least some of the rumors are true. Let's agree that using GPS systems from an early age downgrades some our brain-bound skills at unaided wayfaring, and that believing that something you type will be stored on-device makes you more likely to forget the information than if you are told that after you type it the record will be erased. Let's agree that brain-based recall of a good search string (enabling easy future search) is now often prioritized over on-board storage of the target information itself[22]. Such consequences are intrinsically bad (representing real losses) *only* if you start off by identifying your mind and self with the capacities and activities of your bare biological brain.

But what if you were already best understood as an extended mind, a bio-technologically distributed self? From that perspective, what such results display need not be shrinkage and loss so much as the careful husbanding of our own on-board cognitive capital[23]. In that case, the real worries are more practical ones. Perhaps the online storage or GPS signal is fragile or corrupt? Perhaps the information you retrieve is likely to be false or misleading? Perhaps bio-storage would actually, for some specific instance, be cheaper, better, or more reliable than the alternatives?

These, I believe, are exactly the right kinds of question to learn to ask. If I have taken a drug that makes me unable to concentrate, brain-based wayfaring becomes the more fragile option. If I want to recall a 15-digit sequence after one exposure, I might not trust my biological memory. But I may also estimate that bio-storage remains the best and cheapest option for briefly recalling some much shorter sequence. Quite generally, much of human intelligence involves what are often seen as metacognitive skills – skills of knowing what to rely upon and when. These skills are crucial in cases where we have to decide to what extent to rely on various forms of 'cognitive offloading' rather than brain-based means of storage and recall[24].

Such skill sets now need to be expanded to include, for example, the assessment of suggestions made by off-brain resources such as ChatGPT. We now need to become experts at estimating the likely reliability of a response given both the subject matter and our level of skill at orchestrating a series of prompts. We must also learn to adjust our levels of trust according to our choice of portals and wrap-arounds. Some of these, like FunSearch, can already help mitigate at least some of the risks.

Learning these fluid meta-skills will, to some extent, happen naturally over time. But we should also take steps, both as individuals and societies, to better educate ourselves in roughly how such systems work, hence their characteristic patterns of success and failure. We should also offer practical training in (to take just the most obvious example) how to formulate prompts that ameliorate some of the risks. For example, simply telling ChatGPT to 'answer as if you are a careful, well-informed, academic specializing in such-and-such' goes a surprisingly long way towards nudging its outputs in the direction of measured, evidence-based response.

None of this is to downgrade the very real worries that these powerful new tools suggest. For example, understanding the creative process itself as altered and distributed by the advent of powerful machine intelligences drawing upon huge bodies of human endeavor has legal and economic consequences, and notions of ownership and reward need to respond to these changes[25]. But over time, and with due care and attention, all our stunning new resources should come together, creating a collaborative web that acts as a massive amplifier and transformer of creative human intelligence.

Nested in such a web, human creativity need not be dampened. Instead, it can flourish, continuously targeting new tasks and horizons. All that alters (just as it has always altered) will be some of the specific tasks and processing that human brains perform. These alterations will reflect what can safely be devolved entirely to the new digital unconscious, what remains apt for storage in biological memory, and what must emerge through new kinds of epistemically well-regulated interaction.

This is where we – both as individuals and as societies – have real work to do. As societies, we need to prioritize (and perhaps legislate for) technologies that enable safe synergistic collaborations with our new suites of intelligent and semi-intelligent resources. As individuals, we need to become better estimators of what to trust and when. That means educating ourselves in new ways, learning how to get the best from our generative AIs, and fostering the core meta-skills (aided and abetted by the use of new personalized tools) that help sort the digital wheat from the chaff.

As part of this process, deep and abiding concerns for 'extended cognitive hygiene' will need to be instilled in us from quite an early age. Some of this is happening already. Younger generations are savvier than ever about privacy, phishing, and what to share online. We should now apply equally demanding standards to everything that we might be tempted uncritically to incorporate into our new digitally extended minds. That means developing and applying a rich epistemology (theory of knowledge) – one better suited to the unique sets of opportunities and challenges that confront our bio-technological hybrid minds[26].

Andy Clark
University of Sussex, Brighton, UK. e-mail: andy.clark@sussex.ac.uk

Received: 27 February 2020; Accepted: 5 May 2025;
Published online: 19 May 2025

References

1. Dahmani, L. & Bohbot, V. D. Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Sci. Rep.* **10**, 6310 (2020).
2. Fisher, M. et al. Searching for explanations: how the internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* **144**, 674 (2015).
3. Loh, K. K. & Kanai, R. Higher media multi-tasking activity is associated with smaller gray-matter density in the anterior cingulate cortex. *PLoS One* **24**, e106698 (2014).
4. Barr, N., Pennycook, G., Stolz, J. A. & Fugelsang, J. A. The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Comput. Hum. Behav.* **48**, 473–480 (2015).
5. Firth, J. A., Torous, J. & Firth, J. Exploring the Impact of Internet Use on Memory and Attention Processes. *Int J. Environ. Res Public Health* **17**, 9481 (2020).
6. Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I. & Pechenkina, E. Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *Int. J. Manag. Educ.* **21**, 1472–8117 (2023).
7. Clark, A. & Chalmers, D. The extended mind. *Analysis* **58**, 7–19 (1998).
8. Clark, A. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence* (Oxford University Press, NY, 2003).
9. Clark, A. *Supersizing the Mind: Action, Embodiment, and Cognitive Extension*. (Oxford University Press, NY, 2008).
10. Ward, A. F. Supernormal: how the internet is changing our memories and our minds. *Psychol. Inq.* **24**, 341–348 (2013).
11. Wegner, D. M. & Ward, A. F. How Google is changing your brain. *Sci. Am.* **309**, 58–61 (2013).
12. Ferguson, A. M., McLean, D. & Risko, E. F. Answers at your fingertips: access to the internet influences willingness to answer questions. *Conscious Cogn.* **37**, 91–102 (2015).
13. Goldin-Meadow, S. *Hearing Gesture: How Our Hands Help Us Think* (Harvard University Press, Cambridge MA, 2003).
14. Parr, T., Pezzulo, G. & Friston, K. J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior* (MIT Press, Cambridge MA, (2022).
15. Clark, A. *The Experience Machine: How Our Minds Predict and Shape Reality* (Penguin Random House, NY, (2023).
16. Clark, A. Extending the predictive mind. *Australas. J. Philos.* **102**, 119–130 (2022).
17. Constant, A. et al. A computational approach to selective attention in embodied approaches to cognitive archaeology. *J. Royal Soc, Interface*, **21** 20240508 (epub2024).
18. Shin, M., Kim, J., van Opheusden, B. & Griffiths, T. L. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proc. Natl Acad, Sci.*, **120**, e2214840120 (epub 2023).
19. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
20. Brinkmann, L. et al. Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
21. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
22. Sparrow, B., Liu, J. & Wegner, D. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778 (2007).
23. Hamilton, K. A. & Benjamin, A. S. The human-machine extended organism: new roles and responsibilities of human cognition in a digital ecology. *J. Appl. Res. Mem. Cognit.* **8**, 40–45 (2019).
24. Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C. & Tsai, P. C. Outsourcing memory to external tools: a review of 'Intention Offloading. *Psychon. Bull. Rev.* **30**, 60–76 (2023).
25. Epstein, Z. et al. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
26. Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O. & Pritchard, D. (Eds.) *Extended epistemology* (Oxford University Press, Oxford, 2018).

Acknowledgements

The author was supported by the European Research Council, Synergy Grant (XSCAPE) ERC-2020-SyG 951631.

Author contributions

Andy Clark conceived and wrote the article.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Andy Clark.

Peer review information *Nature Communications* thanks Peter Reiner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025