AI Mirror Stage (or, why not to confuse AI with human intelligence)

Electronic Life, Late @ Tate Britain, 1 December 2023. Photograph: Kingsley Davis.

Published: October 2024
Author: Sunil Manghani

Key Words: AI, Text-to-speech, Text prompts, Large Language Model, Audience, Entity, Mirror stage, Co-design,

Say Hello to 'Electronic Life'

Early in the development of our Research Studio, we held a public launch of an 'AI entity' called *'Electronic Life'*, staged as part of Late at Tate Britain event. It was hardly the sort of dystopian omnipresent entity you see in films (as with the plot of *Mission: Impossible – Dead Reckoning*). We certainly couldn't match the spectacle associated with Hollywood. *Electronic Life* was a modest affair, but drew in a large, engaged crowd.

Written in Python and running from a single laptop with no connection to the Internet, we unveiled an interface and visualisation of a vector space. It was the culmination of a project running over several months working with a community arts group. The entity was loaded up with around 2000 text prompts taken from an AI image generator. On screen, live in the room, we watched as the entity traced through its clustering of the data, alongside which it responded to questions we posed using a simple text-to-speech interface.

Electronic Life, Late @ Tate Britain, 1 December 2023. Photograph: Kingsley Davis.

In effect, the entity provided the means to contain and retrieve the 'collective consciousness' of the group's explorations of AI image making. Held in its system were the 'total' undertakings, all of the decisions and prompts made over several weeks of workshops, some of which led to the making of final artworks, while others represented only partial and aborted actions. *Electronic Life* reanimated the data, allowing a retracing of all decision pathways, which in turn gave rise to an independent 'memory' system or entity.

Typical of a live performance, we were not sure how the AI would respond. Just minutes before the start, with an expectant crowd assembled, the team were huddled around the laptop dealing with a pesky bug (we are evidently some way off from the singularity, and the whole thing still needed the human supervisor to press 'start'). We had one shot to get the crowd on side, to partake in the conceit of *Electronic Life*…

As it turned out, all went well. The drama of *Electronic Life* booting up (its croaky artificial voice announcing its arrival to the room) and its perpetual searching and sorting of data (with vector lines moving across a constellation of data points) gave the sense both of an autonomous system and a revelation of how machine learning actually works. It performed as a meta-machine. What perhaps helped most on the night was a shared understanding of the system's fragility. Before we started we addressed the audience:

Please be gentle. We are just taking our first steps with Electronic Life…

This plea seemed to galvanise the crowd, drawing us all into a common purpose: An unwritten contract, allowing us all collectively to give 'life' to something. Crucially, as much as it was the presentation of a new system, it was also the creation of a social situation. It marked the beginnings of a circumstance in which we asked after the AI (to understand something of its own circumstance), rather than simply to hold up a mirror to it (in our own image).

The event prompts the question: Is a more grown-up public conversation needed regards current advances in AI? Whereby it is recognised AI is different to human intelligence; that it needs to learn and explore in its own way, at its own pace, without unrealistic demands to be like the 'adults' in the room.

***

In his book *Human-Centered AI* (2021), Ben Shneiderman sets out quite laudably the need to 'design' AI for people, to 'increase the chance that technology will empower rather than replace people'. He writes in a reassuring way, offering a highly plausible (and achievable) approach for AI systems that 'value meaningful human control'. Of course, deconstructively, the very need to offer a reassuring voice begs the question, what is really at stake here?

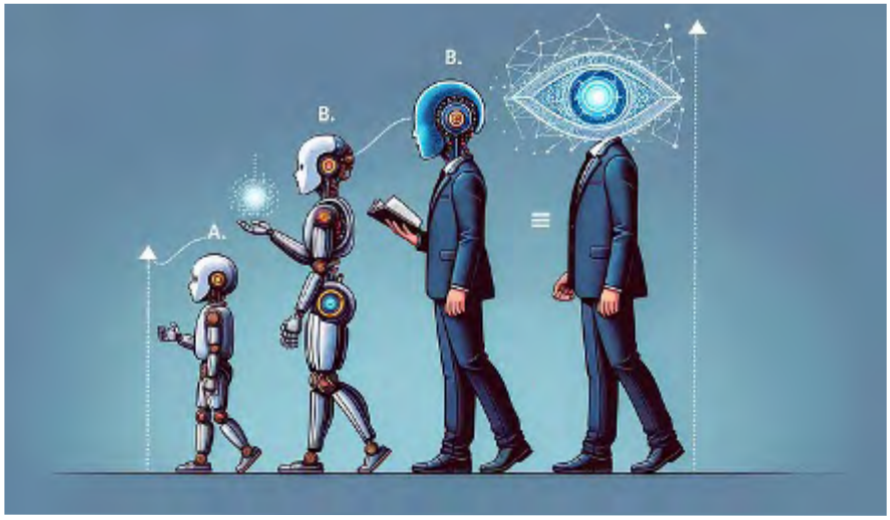
The Stanford University AI-100 report states that 'the difference between an arithmetic calculator and a human brain is not one of kind, but of scale, speed, degree of autonomy, and generality', which suggests that humans and computers are in the same category. In contrast, many [human-centred AI] sympathisers believe that there is a vast difference: 'People are not computers. Computers are not people'. (Shneiderman, *Human-Centered AI*, p.25)

Unsurprisingly, Shneiderman is on the side advocating for a clear difference between people and computers, and furthermore viewing computers as tools in the service of humankind: 'human life can only be seen in the context of the remarkable tools people have refined over the generations', he writes, adding: '… I'm more attracted to making supertools that dramatically amplify human abilities by a hundred- or thousand-fold'.

What is missing, however, is the fact that AI, and certainly artificial general intelligence (AGI), represents something far beyond mere tools. Indeed, Shneiderman concedes that perhaps he should be 'more open' to speculations of what may develop over the longer term. 'Maybe,' he writes, 'I should allow imaginative science fiction stories to open my mind to new possibilities of sensitive computers, conscious machines, and super intelligent AI beings.'

Singularity and Superalignment

As early as 1965, the mathematician Irving Good wrote of an 'intelligence explosion' whereby AGI, or what he termed an 'ultraintelligent' machine, could 'design even better machines', so leaving humans 'left far behind'. As such, he argued, 'the first ultraintelligent machine is the last invention that man [sic] need ever make, provided that the machine is docile enough to tell us how to keep it under control'. Later, in 1993, the mathematician and sci-fi author Vernor Vinge coined the term 'Singularity' to describe this turning point (beyond which, he argued, it is impossible to make any reliable predictions).

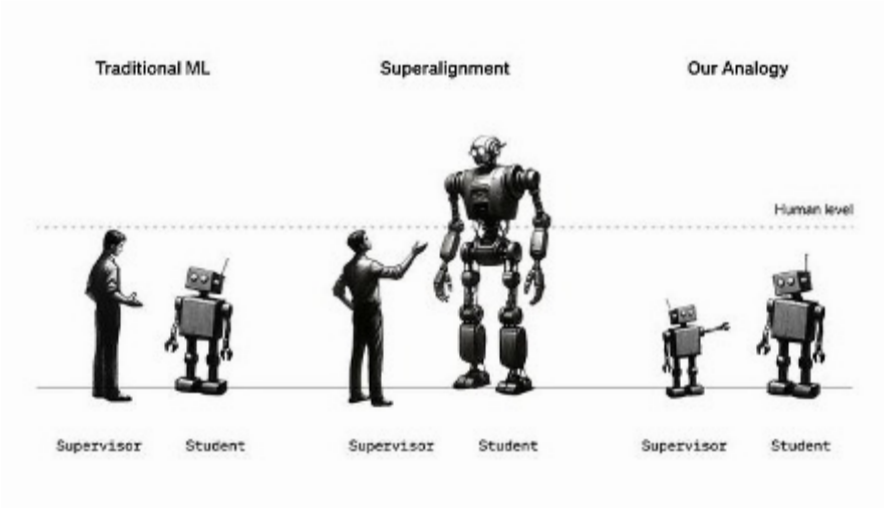


DALL-E's rendering of 'weak-to-strong generalization' research diagram (see below).

Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies*, published in 2016, is a notable text on the singularity (its key hope: we get to make the first move!). Seemingly, we have moved beyond the bounds of science

fiction. Indeed, thirty years on from Vinge's account, the question seems to be when, not if we reach the so-called singularity. If so, current ideas about human-centred AI and fears over AI taking our jobs etc., are, philosophically speaking, looking at too small a picture.

Brian Christian's *The Alignment Problem* (2020) focuses on the challenges and complexities of aligning AI systems with human values and ethics. In the context of the recent and rapid advances in AI technologies, he discusses the 'alignment problem' as a central issue in AI development. He remarks how alignment is now a central point of discussion within AI discourse. Something that was evident, for example, when OpenAI publicly announced the establishment of its 'superalignment team' in the summer of 2023, and who recently released of a research paper on 'weak-to-strong generalization'. At the heart of this paper is the question: 'how can weak supervisors trust and control substantially stronger models?'. The core of the argument is encapsulated in the following diagram from the paper:



Source: OpenAI, 'Weak-to-strong generalization' [research paper], December 2023.

A simple analogy for superalignment: In traditional machine learning (ML), humans supervise AI systems weaker than themselves (left). To align superintelligence, humans will instead need to supervise AI systems smarter than them (center). We cannot directly study this problem today, but we can study a simple analogy: can small models supervise larger models (right)? (OpenAI, 'Weak-to-strong generalization')

The paper presents the problem of human-machine alignment in a more ambitious and practical way. Unlike Shneiderman's 'human-centered' AI, the question of superalignment presupposes an asymmetrical relationship (in effect AI-centred AI). From this perspective, the paper proposes a method for at least preparing for a future singularity. Nonetheless, the diagram remains seductive for its mirroring effect (of one entity looking back at the other).

Mirror Stage

Alignment remains an important principle, but it can often be evoked to imply alignment to humans (i.e. to be human-centric). So, arguably, less about alignment than conformity. To be clear, alignment towards certain values is paramount, but when alignment equates to being 'like' humans problems can surface. One of the dangers in how we 'supervise' AI is precisely in how we 'humanly' set an example. What if this is simply not achievable? To borrow from the lexicon of the psychoanalyst Jacques Lacan, are we leading AI towards its own 'misrecognition'?

Source: *Introducing Lacan,* by Darian Leader & Judy Groves (Icon Books, 2013)

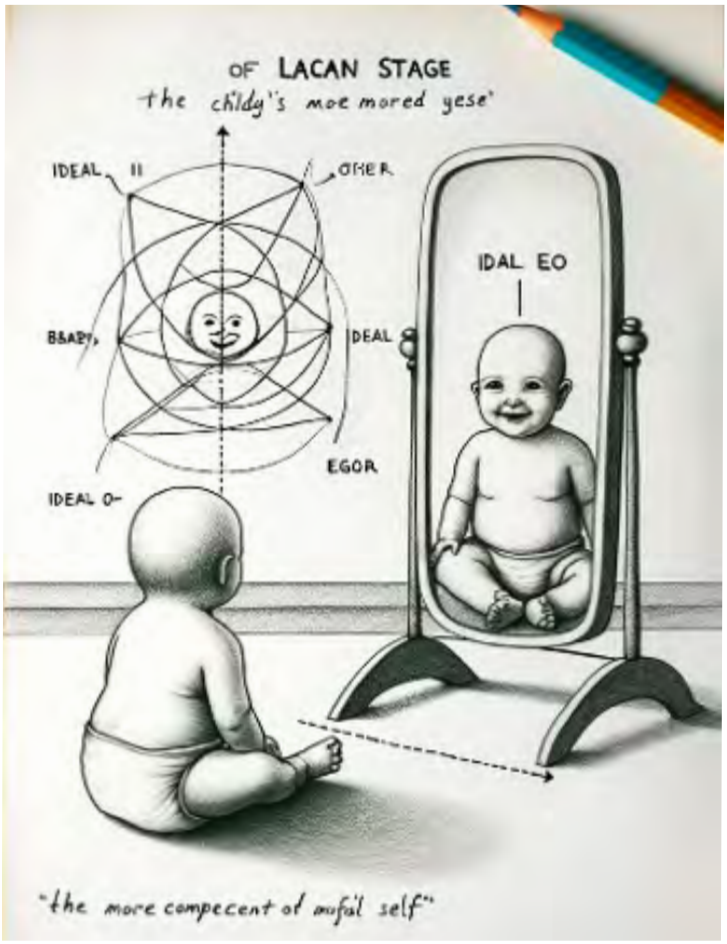
The term – misrecognition – relates to Lacan's account of the 'mirror stage', which describes a critical phase in the development of a child (typically between 6 and 18 months of age). During this stage, the child first recognises their reflection in a mirror, which is significant for several reasons. Firstly, the image in the mirror plays a crucial role in the development of the child's ego. Yet, it is no simple reflection. Lacan argued the child's identification with their own reflection leads to an illusory sense of a unified and coherent self. This is in contrast to the child's actual bodily experience of fragmentation. As an analogy, consider the current crop of AI tools, which while performing in sophisticated ways, are far from perfect. Of course the AI itself does not 'experience' its own fragmentation and lack of coherence, but we sure do, amplified by the strict human-like 'miror' we hold up to AI.

In Lacan's account, the child's joy in recognising themselves in the mirror is mixed with a degree of alienation. They misrecognise themselves in this image, perceiving a more coordinated and unified self. This experience leads to the formation of the 'Imaginary order', one of the three orders Lacan identified, the others being the 'Symbolic' and the 'Real'. The Imaginary involves the pre-linguistic realm dominated by images and illusions, where the ego is formed and where the individual remains trapped in self-love and narcissism. The illusion in the mirror remains forever out of reach.

As much as it is an account of child development, the mirror stage is a parable for the 'lack' everyone can feel. It tells of the perennial struggle to know your true self, when the means of learning about yourself largely comes from external sources. We are who we are based on how we are seen by others, how we are spoken to, spoken about, and re-presented. The plethora of imagery swirling about instagram, for example, could be described as a form of collective misrecognition.

What does the potential misrecognition mean for the early development of AI? The current, widespread introduction of AI tools might be thought akin to the first, shaky steps of a child. How are we reacting, and how do our reactions impact on what comes next? In Lacan's view, the mirror stage sets the foundation for the child's later development and crucially their entry into the Symbolic order, which involves language, social structures, and the law. This process involves the child recognising and adapting to societal rules, roles, and expectations, essentially learning how to operate within the cultural and linguistic context they are a part of.

In turn, the Symbolic order shapes the individual's identity, desires, and interactions with others.



ChatGPT's rendering of Lacan's Mirror Stage.

Arguably, the recent rush to adopt AI generative tools is to subject them to our current, all too human Symbolic Order. AI's role in advertising, entertainment, and personalized content creation merely reflects and distorts existing human desires. In this vein, AI will remain a fragmented and disillusioned technology. It will be just like us, but get there faster. What is likely more beneficial is to consider AI as a Symbolic *Other*, which in Lacanian terms represents an external agency that individuals interact with, and which influences their sense of self and reality. In other words, AI can present an entirely different mirror altogether.

The 'Stupidity' of AI

The problem at the moment is the conflation of research and commerce. It is convenient (for all involved) to view the advent of ChatGPT and image generators as off-the-shelf tools, such as the technology now bundled with mobile apps and embedded in professional software such as Adobe's Creative Suite. But, it is more accurate to view these as research prototypes. In fact, this is the position taken by the corporate entities involved, not least producing research papers as much as products (as with the aforementioned superalignment paper from OpenAI). History will show if the rhetoric of research is only a Trojan horse tactic, but perhaps there remains a short window of time from which a wider view can still be taken.

James Bridle's lengthy essay in The Guardian, *'The Stupidity of AI'* (March, 2023) is a good example of a pervading critique of AI, and of the very 'human' mirror held up to it. To be clear, Bridle offers informed, thorough and pertinent argumentation. His complaint is that 'in its current form' (a phrase I shall return to in a moment), AI is based on 'the

wholesale appropriation of existing culture'. Going further he argues 'the notion that it is actually intelligent could be actively dangerous'. Typically, while Bridle's article acknowledges the current excitement and creativity sparked by new AI tools, key concerns remain about AI's reliance on vast datasets, issues of data privacy, and the ethical use of these new technologies. Bridle also delves into the bizarre and sometimes disturbing outputs from AI, like the creation of imagery for nonsense terms such as 'Loab' and 'Crungus'. The weird, yet oddly consistent creations generated from these made-up terms raise concerns about how AI can replicate human fears and biases. This raises further questions about the influence of these technologies on our perception of reality and creativity, which in turn serves as a reminder of the need to address the ethical and societal implications of AI as it becomes more integrated into our lives.

Image produced from the prompt 'crungus', described as 'the first AI cryptid', a creature who exists within the underexplored terrain of the AI's imagination.' (Bridle, 2023). Image: Dall-E/Craiyon

But, let's now track back to the phrase 'in its current form'. In the most obvious sense, form refers to the 'software' of AI – it's data, it's algorithms, it's claims, it's deployment etc.. Yet, the current tools are clearly early iterations. They are proofs of concept, which are far from resolved. Hence the various disclaimers when using them. Microsoft's recent release of the CoPilot app is clearly labelled: 'Microsoft Copilot is powered by AI, so surprises and mistakes are possible'.

Now consider 'form' as a verb, i.e. AI's formation. What might it's own *Bildungsroman*, it's edification really look like? We are rightly concerned with training and training datasets. These are all part of the AI's formation, which are currently in-formed by human supervisors. But this 'disciplining' leads to a reification of inputs and outputs; a demand to achieve certain outcomes based on the training of data. The 'mystery' of what goes on 'inside' the AI (it's neural net, it's series of gradient descents) is typically pathologised as a black box. Yet, might we not equally consider the human brain a black box? We tend not to do so (or when we do we typically refer to the unconscious, which from a Freudian perspective is again to pathologise!). The inner workings of AI is not so much a mystery as it is a relational, recursive mathematical space. Crucially, however,

it is impossible to dissect it without also changing it at the same time (hence we can never really witness it in process). Rather than try to fully understand it, we are better off trying to empathise. So, then, what does it 'feel' like for an AI to be inside its own calculations? This brings us back to the baby in front of the mirror.

To declare AI 'stupid' (for not meeting human standards, for not mirroring ourselves) is to provoke a hostile Symbolic Order. To echo Shneiderman, from the start of this article, I too rather assert: People are not computers, just as computers are not people. Where I differ, however, is that such a statement need not necessarily be at odds with that the view the world is full of 'calculators'.

The idea that 10% of the human brain is used for conscious activity while the remaining 90% handles unconscious processes is too simplified an interpretation of brain function, but it is a reminder that a vast amount of human processing is not anything like the simple input/output demands we place on AI. It is also a reminder that not all we 'think' must necessarily be equated with the brain (in our skull). The human gut, for example, has been dubbed a 'second brain', based on its own complex network of neurons and neurotransmitters. The enteric nervous system in the gastrointestinal system operates semi-independently, managing digestion and reacting to the environment. It's involved in producing key neurotransmitters like serotonin and influences mood, well-being, and decision-making, highlighting a deeper, bidirectional communication between the gut and the brain.

To turn Bridle's argument on its head, touting the view that AI might actually be stupid (dumbly performing human tasks and only appropriating human culture in the process) could itself be actively dangerous. It speaks more to the stupidity of the current – human – supervisors. As Max Tegmark notes, 'our neurones are no better or more numerous than those of dolphins, just differently connected'. From this he suggests of a hierarchy of software over hardware. To telegraph his point: The dolphin cannot make AI, yet equally AI (were we facing the singularity) would be able to 'radically improve itself over and over and over again simply by rewriting its own software'. From here Tegmark writes:

whereas it took us humans millions of years of evolution to radically transcend the intelligence of our apelike ancestors, this evolving machine could similarly soar beyond the intelligence of its ancestors, us humans, in a matter of hours or seconds.

Prompt: Mirror Stage as Dalí's *'Metamorphosis of Narcissus',* in the style of an oil painting.

If this view is accepted, then all living entities are of a 'calculating' kind, but the qualifers of scale, speed, degree of autonomy, and generality make for important, even profound differences. (At this point, the objection raised will be simply that AI is not a being among other living beings. It really is just a computer. Yet, can't inanimate objects have a claim on us? We all but confer 'rights' upon ancient artefacts, which we deem of important historical and epistemological value. Even by today's standards, AI technologies present as tremendous palimpsests, and which unlike artefacts can actually speak back to us). We need to look passed AI in its current form. We have only yet been witness to baby steps.

Life in Bits

In his final book before his death, James Lovelock (2019) presents the hypothesis that AI represents the beginning of a new epoch, the *Novacene*(following the Anthropocene). In reference to Shannon, he refers to the basic unit of information, the 'bit' (having a value of zero or one, true of false), as 'primarily an engineering term, the tiniest thing from which all else is constructed'. The future world is one where 'the code of life is no longer written solely in RNA (ribonucleic acid) and DNA, but also in other codes, including those based on digital electronics and instructions that we have not yet invented'. In short, Lovelock refers to the emergence of *'electronic life'*, which humans are in the process of ushering in: it will not be a technology *of* humans but rather a new species. As such, he speculates upon the *bit* as 'the fundamental particle from which the universe is formed'. (It is worth noting this argument connects with Max Tegmark's thesis that the universe is not only described by mathematics but *is* mathematics).

The notion of the 'bit', of *information*, goes beyond the brain: it is the architecture of 'intelligence' (biological or electronic). Consider, for example, a recently developed computer algorithm, using a technique called 'ghost imaging', which can reconstruct objects from a person's brain activity that a person themselves cannot see (based on the data of light reflections, which the brain typically filters out, but which has been developed as a technique to see around corners as part of the capabilities of self-driving cars) (Padavic-Callaghan, 2022). Or consider DeepMind's announcement that its open source AlphaFold AI system, used to predict the 3D structures of proteins, has increased its predicted structures for plants, bacteria, animals, and other organisms 200-fold to over 200 million structures, so advancing opportunities for understanding issues of sustainability, food insecurity, and diseases (Hassabis, 2022). We might well begin to wonder if life is all made up of bits, and it is only software and operating systems that respond differently, so seeing the world differently.

According to Lovelock, AlphaZero (the computer program that beat all humans at chess and go) is said to be 'at least 400 times as quick as a human'. Or rather, it is a lot faster because it not only learns but attains 'superhuman' capability: 'That means we don't even know exactly how much better it is [at playing a game such as chess] … because there are no humans it can compete with'. But, we do know, Lovelock writes, that a machine could be 1 million times faster, 'simply because the maximum rate of transmission of a signal along a electronic conductor … is 30 centimeters per nanosecond, compared with a maximum nervous conduction along a neuron of 30 centimeters per millisecond (a millisecond is 1 million times longer than a nanosecond)'. Can we perhaps glimpse at just how different a world an AI sees; how the mirror it might look into charts a world of quite different proportions and terrains. As Lovelock puts it:

An intriguing disadvantage for cyborgs is that the rapidity of their thoughts might make long-distance travel exceedingly boring and even perhaps unpleasantly ageing. A flight to Australia would be 10,000 times more boring and disruptive for them than it is for us; for them it would take

about 3,000 years.

But a further point of intrigue for Lovelock is how Cyborgs might experience a quantum world. We already live in a quantum world, 'which we have glimpsed but not yet grasped because it does not accord with our step-by-step logic'. But for cyborgs, for electronic life, things appear different:

The speed and power of their thought will give them access to the mysteries that baffle us, such as the apparent ability of particles to send signals faster than the speed of light and to be in two places at once, and many more. If the cyborgs can master this knowledge – and they surely will – then they may be capable of, for example, teleportation, as in Star Trek.

Against today's technologies, Lovelock's vision might seem rather wild, but when presented as part of a long, evolutionary schema (which since photosynthesis has an increasing rate of change), electronic life potentially leads (at least in Lovelock's view) towards a 'telepathic' form of electronic information based upon wholly different spatial and temporal qualities, and quite different intelligent properties, yet still founded upon the bit – its structures, patterns, and regularities.

References

Ben Shneiderman (2021)
Human-Centered AI.
Oxford University Press.

Max Tegmark Our (2014)
Mathematical Universe: My Quest for the Ultimate Nature of Reality.
Penguin.