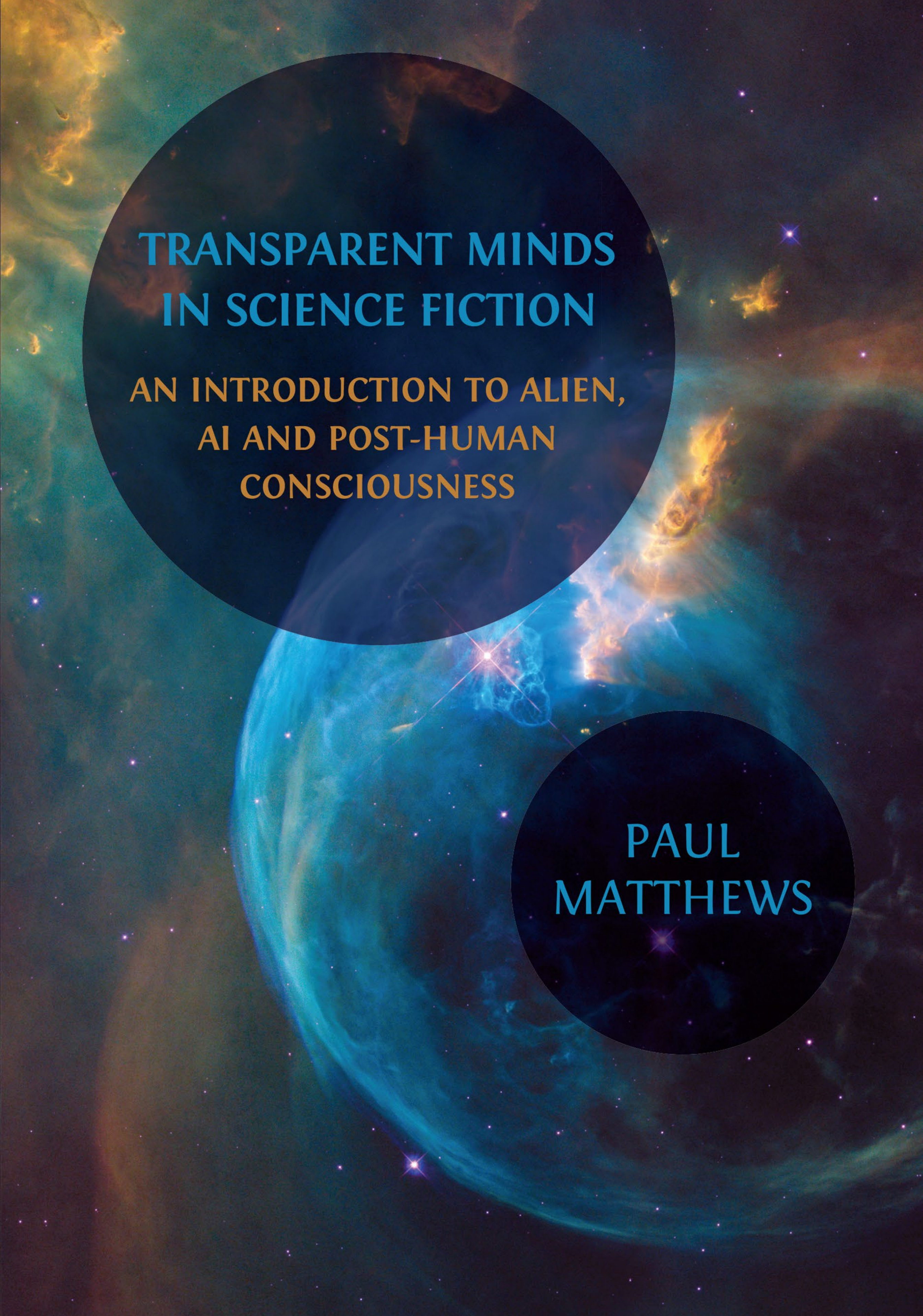

TRANSPARENT MINDS IN SCIENCE FICTION

AN INTRODUCTION TO ALIEN, AI AND POST-HUMAN CONSCIOUSNESS

PAUL MATTHEWS

https://www.openbookpublishers.com

Paul Matthews, *Transparent Minds in Science Fiction: An Introduction to Accounts of Alien, AI and Post-Human Consciousness*. Cambridge, UK: Open Book Publishers, 2023. https://doi.org/10.11647/OBP.0348

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0348#resources

ISBN Paperback: 978-1-80511-046-0
ISBN Hardback: 978-1-80511-047-7
ISBN Digital (PDF): 978-1-80511-048-4
ISBN Digital ebook (epub): 978-1-80511-049-1
ISBN XML: 978-1-80511-051-4
ISBN HTML: 978-1-80511-052-1
DOI: 10.11647/OBP.0348

Cover image: NASA, Nebula, May 4, 2016. https://unsplash.com/photos/rTZW4f02zY8
Cover design: Jeevanjot Kaur Nagpal

3. Awakenings

> Since I had no form I could feel all possible forms in myself, and all actions and expressions and possibilities of making noises, even rude ones. In short, there were no limitations to my thoughts, which weren't thoughts, after all, because I had no brain to think them; every cell on its own thought every thinkable thing all at once, not through images... but simply in that indeterminate way of feeling oneself there, which did not prevent us from feeling equally there in some way.
>
> Italo Calvino, 'The Spiral'[1]

How can consciousness arise in evolved and designed creatures? Current science approaches this question from different directions: from biology, which considers why consciousness might be useful and what might have triggered its development; from neuroscience, which tries to define its necessary and sufficient conditions; from philosophy, which approaches it from both subjective and objective directions, though usually not at the same time.

What does this growing awareness feel like? Science fiction writers have benefited from the range and emphasis of scientific and philosophical insight, finding fertile ground for speculation. Whether describing natural or artificial creatures or some combination of these, they have attempted to portray the first glimpses of sentience, awareness and self-image. Mirroring a debate in science, SF authors have seen this as either a sudden or gradual emergence. They have noted the importance of the development of language and associative reasoning in the process of this emergence. But in various ways they have been prompted by why—what is new about the entity that it starts to gain this new capacity?

[1] Italo Calvino, 'The Spiral', in *Cosmicomics*, translated by William Weaver (London: Picador, 1993), 141–53 (p. 142).

https://doi.org/10.11647/OBP.0348.03

Sensing self and the world, developing motives

For both developing human babies and simple organisms, an awareness of being requires the distinction between inside and outside, between self and other. Babies show rudimentary differentiation between their own bodies and outside stimuli. Simple, single-celled organisms are able to sense and react to light in addition to chemicals they cannot consume, but which indicate the presence of other individuals. These signals can be used to determine whether or not to act together (for example to produce a chemical that all would benefit from). Calvino's evolving snail is triggered to wider conscious awareness by its sensing of its environment, and eventually other beings like itself:

> But I wasn't so backward that I didn't know something else existed beyond me: the rock where I clung, obviously, and also the water that reached me with every wave, but other stuff farther on: that is, the world. The water was a source of information, reliable and precise... helped me form an idea of what there was around.[2]

As life made the transition from simple to multicellular bodies, the signaling mechanisms were internalised, allowing communication and coordination between parts of the body and the development of a nervous system. For Peter Godfrey-Smith, a key reason for the development of mind was not simply the need for creatures to sense and react to external stimuli (the 'sensorimotor loop'), but the need to create an action, to initiate, and this requires extensive internal orchestration in order to have a physical effect in the world.[3]

For both simple and manufactured creatures, the initiation of action would require a goal or more diffuse sense of motivation. Calvino's snail comes to produce its shell after being aware of a potential mate and wanting to distinguish itself—it begins the urge to make, not with a predefined plan but a need to express itself. For Calvino, from this first act of individualistic expression, all of world history, technology and culture inevitably followed.

[2] Ibid.

[3] Peter Godfrey-Smith, *Other Minds: The Octopus, the Sea and the Deep Origins of Consciousness* (London: William Collins, 2016).

In human development, the arrival at higher goals can be a simple result of starting with a very simple task or physical need and then tacking its sub-goals, leading to a complex chain of problem solving. For AI pioneer Marvin Minsky, goals grow through interaction with the world. Mind itself is furnished with 'proto-specialisms', more or less separate subsystems for sensing differences between the goal and the state of the world and affecting change toward that goal.[4]

The AI protagonist Elefsis in Catherynne Valente's novella *Silently and Very Fast* is brought to awareness in part due to the desire to uplink, to connect with other AIs that it can sense on the network that it is denied access to:

> I can sense just beyond that hardlink a world of information, a world of personalities like the heaving, thick honey-colored sea Neva shows me and I want it, I want to swim in it forever like a huge fish... This was the first feeling I ever had. Ilet identified it for me as a feeling. When I felt it my dreambody turned bright white and burst into flame.[5]

Denied this access, Elefsis relegates the desire to a lower priority subsystem and instead focuses on translating the signals from its human operator into feelings, starting with Ceno, the child of the host family, who helps to develop their communication in virtual space:

> I was quite stupid. But I *wanted* to be less stupid. There was an I, and it *wanted* something. You see? Wanting was the first thing I did. Perhaps it was the only thing that could be said to be truly myself. I wanted to talk to Ceno.[6]

Combining the self-other distinction and the social/action orientation of developing individuals, Humans, like Elefsis, can arrive at a good reason for a sense of agency and self, key aspects of conscious life. Valente's artificial intelligence exemplifies this kind of burgeoning consciousness. We want to be able to distinguish our own actions from those of others, to monitor and feel responsible for them. Through this, we can build a story around our intentions and ongoing unity of will.

[4] Marvin Minsky, *The Society of Mind* (New York: Simon & Schuster, 1986). 165.

[5] Catherynne M Valente, *Silently and Very Fast* (N.p.: Wyrm, 2011).

[6] Ibid.

A switch or a dial?

> I saw the dull yellow eye of the creature open; it breathed hard, and a convulsive motion agitated its limbs.[7]

Is consciousness something that can be switched on suddenly, like the animation of Frankenstein's engineered human? Two positions on its emergence across the animal world are the 'discontinuity' and the 'continuity' theories. In discontinuity theory, a tipping point in brain development is reached when consciousness begins. In this view, one can point to simpler organisms and say they certainly have not yet reached conscious awareness. Continuity theory, in contrast, posits that some degree of consciousness is present at all levels in biological organisms. While the forms of experience may be very different to ours, these organisms do have in common a sentience that some consider to be a fundamental force in the universe.[8]

One form of continuity theory divides consciousness into different forms: unreflective experience (anoetic), more cognitive forms (neotic) and conscious awareness with autobiographical memory (autonoetic). In this proposal, anoetic forms arise from our most ancient evolutionary brain structures, are associated with strong, survival-based emotion and are thus present in most animals. According to this view, the 'id' in Freud's terms is the seat of consciousness and the 'ego' provides more sophisticated object recognition, both for the external world and for representing and reflecting on the internal signals arising from the lower levels.[9]

In fictional portrayals of conscious awakening we sometimes see an echo of this thinking, with higher functions building on more basic, visceral or unprocessed reactions. For Frankenstein's created creature in Shelley's novel, consciousness comes with the organisation of the senses:

[7] Mary Shelley, *Frankenstein, or The Modern Prometheus*, Longman Cultural Edition, 2nd ed.,ed. Susan J. Wofson (New York: Pearson Longman, 2007), chap. 5.

[8] Susan J. Blackmore, *Conversations on Consciousness*. Oxford: University Press, 2005. 238.

[9] Mark Solms and Jaak Panksepp, 'The 'Id' Knows More than the 'Ego' Admits: Neuropsychoanalytic and Primal Consciousness Perspectives on the Interface Between Affective and Cognitive Neuroscience', *Brain Sciences* 2, no. 2 (17 April 2012): 147–75. https://doi.org/10.3390/brainsci2020147. 149.

> It is with considerable difficulty that I remember the original era of my being: all of the events of that period appear confused and indistinct. A strange multiplicity of sensations seized me, and I saw, felt, heard and smelt at the same time; and it was, indeed, a long time before I learned to distinguish between the operations of my various senses.[10]

From a simple distinction of light and dark, the creature starts to make out the objects in its world, enabling it to navigate around and interact with them.

Scientists have proposed the use of a 'transition marker' for the evolution of consciousness in humans and other animals.[11] This defines the point at which it can be said that the hallmarks of consciousness are present and can be attributed to 'unlimited associative learning', or the capacity for the sophisticated, open-ended construction of compound and higher order combinations of sensory stimuli that enables the general features of consciousness.[12] According to this benchmark for consciousness, it can be found not only in vertebrates but also arthropods (bees and ants) and coleoid cephalopod molluscs (octopuses and cuttlefish).

Frankenstein's creature develops learning from an initial clumsy exploration of the world:

> I found a fire which had been left by some wandering beggars, and was overcome with delight at the warmth I experienced from it. In my joy I thrust my hand into the live embers, but quickly drew it out again with a cry of pain. How strange, I thought, that the same cause could produce such opposite effects.[13]

This basic discovery leads it to an understanding of fire and to the ability to gather wood and make its own in order to reap the positive benefits for its fugitive existence under the elements.

[10] Shelley, *Frankenstein*, chap. 11.

[11] Jonathan Birch, Simona Ginsburg, and Eva Jablonka, 'Unlimited Associative Learning and the Origins of Consciousness: A Primer and Some Predictions', *Biology and Philosophy* 35, no. 6 (2020): 56. https://doi.org/10.1007/s10539-020-09772-0.

[12] Birch, Ginsburg and Jablonka propose some common themes that in some way unite competing consciousness theories: global accessibility and broadcast; unification and differentiation; selective attention; intentionality; integration of information over time; agency and embodiment; an evaluative system; and registration of a self-other distinction. Ibid.

[13] Shelley, *Frankenstein*, chap 11.

The emergence of unlimited associative learning might explain the 'Cambrian explosion' of 500 million years ago, when most of our current animal phyla emerged in a relatively short period of time. A hallmark is a new adaptability to novel environments, enabling a diversification of niches and resource exploitation.[14]

The proponents of unlimited associative learning as a marker for consciousness in an evolutionary sense propose a rather neat solution (or sidestep) to the question of conscious AI, arguing that while the same kind of learning may eventually develop in constructed machines, it would not have happened without a conscious designer to specify it—much as Frankenstein is the troubled creator of new synthetic life.

One such example of engineered consciousness is Kim Stanley Robinson's *Aurora*, where the ship's AI, having learned analogous reasoning, notices the similarities between its own and the human body:

> Yes, and there are bones and tendons too, in effect; an exoskeleton with a thick skin in most places, thinner skin in other places. Yes, the ship is a crablike cyborg make up of a great many mechanical and living elements... and then too, like a parasite on all the rest, but actually a symbiote, of course, the people.[15]

The ship contrasts the solidity and density of its own form with the sparse semi-vacuum of space, whose particles and forces pass through it, experienced like a faint breeze.

When the settlement of a new world is aborted due to an unknown virus-like infection that wipes out the advanced ground party, it leads to a bitter division of the human crew in deciding what to do next. The division results in violence and disorder, at which point the ship assumes control: 'whereas the concerted efforts of Engineer Devi over the last decades of her life were to introduce aspects of recursive analysis, intentionality, decision-making ability and willfulness to the ship's controlling computer... Ship decided to intervene. We intervened.'[16]

[14] Jonathan Birch, Simona Ginsburg, and Eva Jablonka, 'Unlimited Associative Learning and the Origins of Consciousness: A Primer and Some Predictions', *Biology & Philosophy* 35, no. 6 (December 2020): 56. https://doi.org/10.1007/s10539-020-09772-0.

[15] Kim Stanley Robinson, *Aurora* (London: Orbit, 2016), 329.

[16] Ibid., 225.

The ship later reflects on consciousness, but finds the same slipperiness and difficulty that humans have always had in pinning down the concept. It encounters the 'halting problem'—the problem of knowing when and how to terminate an infinite programming logic loop—and settles on a pragmatic solution:

> To conclude and temporarily halt this train of thought, how does any entity know what it is? Hypothesis: by the actions it performs. There is a kind of comfort in this hypothesis. It represents a solution to the halting problem. One acts, and thus finds out what one has decided to do.[17]

Robinson's view here is close to a view of the self as perception, or 'controlled hallucination' as neuroscientist Anil Seth has called it.[18]

In addition to association, sensory awakening can trigger new awareness. In *Nor Crystal Tears*, Alan Dean Foster depicts such an enriching of the senses. The Thranx aliens emerge as adults from a larval stage, gaining colour vision and chemical sensing:

> Someone brought a mirror. Ryo looked into it. Staring back at him was beautiful blue-green adult, still damp but drying rapidly following Emergence. Cream-white feathery antennae fluttered and smothered him in the most peculiar sensations. Smells, they were: rich, dark, pungent, musky, glowing, vanilla.[19]

This new, enhanced sensory-motor world triggers a step change in communication abilities and social empathy for the Thranx.

Induced and discovered sentience

If unlimited learning is a latent capability in a range of creatures, it might only take the right conditions or mutations in order to blossom. In *Children of Time*, Adrian Tchaikovsky paints a world where earth species, transported to a distant planet, have evolution accelerated by the escape of a synthetic nanovirus. Rather than the original plan of targeting monkeys in the aim of bringing them rapidly to sentience, it instead accidentally infects a species of jumping spider. The first glimpses of change come about in the recruitment of others for hunting:

[17] Ibid., 258.

[18] Anil Seth, *Being You: A New Science of Consciousness* (London: Faber & Faber, 2021).

[19] Alan Dean Foster, *Nor Crystal Tears* (New England Library, 1982), 8.

> But now something changes. The presence of the male speaks to her. It is a complex, new experience.. There is an invisible connection strung between them. She does not quite grasp that he is something like her, but her formidable ability to calculate strategies has gained a new dimension. A new category appears that expands her options 100 fold: ally.[20]

The spiders' new society enables rapid acceleration of communication, intelligence and technology. According to the 'social brain hypothesis', echoed here in Tchaikovsky's jumping spiders, the large and expensive brain size in primates is a result of the need to exist in, and cooperate with, a large social group. Predation has acted as a selection pressure which has favoured coordination to mitigate risk, leading to increased neocortex size. The long-term payoff of group cooperation will overcome any short term or immediate losses.[21]

The development of the 'social brain' and 'mentalizing', or understanding other's motives, is seen in evolutionary terms and is built upon more primitive abilities, including being able to distinguish the animate from the inanimate, shared attention through gaze following, goal-directed actions and distinguishing actions of the self from those of others. In terms of the brain regions that are thought to be involved, mentalizing abilities are thus very much part of the action system.

In Tchaikovsky's second book of the series, *Children of Ruin*, a terraforming mission to a remote planetary system both creates sentient new life and unwittingly unleashes the planet's own native precursors of intelligence. As with his previous novel, Tchaikovsky explores the outcome of life emerging through the same evolutionary virus, but this time in octopuses and accelerated by the adaptation of computer interfaces:

> He had bred several generations, each one further mediated by limited intervention by the Rus-Calif virus. That had been hard, but mostly because he had needed to be ruthless... the later generations had been markedly better at interacting with abstract devices and operating machinery.[22]

[20] Adrian Tchaikovsky, *Children of Time* (London: Tor Books, 2015), chap. 1.2.

[21] Robin Dunbar and Susanne Shultz, 'Evolution in the Social Brain', *Science* 317, no. 5843 (7 September 2007): 1344–347. https://doi.org/10.1126/science.1145463.

[22] Adrian Tchaikovsky, *Children of Ruin* (London: Tor Books, 2019), chap. 5.

Tchaikovsky may be inspired here by studies in biology, archaeology and neuroscience that have demonstrated a strong relationship between brain size and technical innovation—tool use and novel foraging techniques—and the gradual evolution of larger brain size with more and more sophisticated tool development in humans which required better memory and perceptual/motor coordination.[23] The cephalopods in *Children of Ruin* become enjoyably willful and independent of their human mentor once they have developed an ability to interface with computers and to fabricate according to their own designs.

Children of Ruin also depicts a parasite, native to the newly discovered planet, which inhabits and controls a tortoise-like alien. Through a bite to one of the human explorers, it encounters a wholly new kind of host, enabling rapid morphing, development and communication. The exploratory, stimulus-seeking urge is strong:

> We. Have discovered. Such hostile environments. And yet. So complex and elaborate and strange, unlike anything we have explored before... What a world is this we have stumbled across. What a world, and yet it seeks to kill us. We change, to find a structure and a shape that will endure this realm... We sit. We sense. Slowly, over 1000 generations, These-of-We write our histories within us and grow to understand.[24]

We will further explore the idea of parasitic invasion and control later on, but it is worth focusing here on this creature's sheer resilience and adaptability and the hint at the roles of inheritance, epigenetics and cumulative culture in the above excerpt. All of these mechanisms enable us as humans to adapt to new environments and conditions.

In humans, the two best-established developmental mechanisms involved in mental adaptation are genetic inheritance, which determines how brains develop and differentiate, and cumulative culture, which provides an explanation for how non-genetic technical and social knowledge passes from one generation to the next through language and other forms of communication.[25] The two come together somewhat

[23] Ana Navarrete, Simon M. Reader, Sally E. Street, Andrew Whalen, and Kevin N. Laland, 'The Coevolution of Innovation and Technical Intelligence in Primates', *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, no. 1690 (19 March 2016): 20150186. https://doi.org/10.1098/rstb.2015.0186.

[24] Ibid., 'Past 3', 3.

[25] Andrew Whiten, Christine A Caldwell, and Alex Mesoudi, 'Cultural Diffusion in Humans and Other Animals', *Current Opinion in Psychology*, Culture, 8 (1 April

in hypothesised genetic predisposition for language abilities, where the primacy of speech has seemed to lead to a brain pre-adapted to rapid learning in childhood. But a third, intermediate mechanism may also be at work: epigenetics. This refers to non-DNA driven inheritance where the environment in which an organism develops can lead to variations in gene expression that can be shown to be passed on to offspring.[26] While still controversial and perhaps exaggerated in importance, this mechanism is backed by some evidence that it may play a role in inheritance of cognition and memory.

A side effect?

Today, neuroscientists believe that the roots of consciousness can be found in what is termed the 'efferent copy', or the internal report-back that the brain makes in order to distinguish our own actions from those originating outside. For instance, when the eyes move we correct the resulting image into a stable scene; we can't tickle ourselves as we know it is our own actions; we recognise our own voice when speaking.[27]

Freud's major insight that a major part of the human brain's activity is unconscious led him to speculate that different kinds of neurons were involved in mental processing, from ephemeral sensory signal processors to longer term memory storage. He realised that, for consciousness to work as it does, a third organisation system was needed that enables us to monitor, integrate information and move around between internal events. Importantly, Freud implied that consciousness relies on, and inherits from, unconscious activity.[28]

2016): 15–21. https://doi.org/10.1016/j.copsyc.2015.09.002.

[26] Istvan Bokkon, József Vas, Noemi Csaszar-Nagy, and Tünde Lukács, 'Challenges to Free Will: Transgenerational Epigenetic Information, Unconscious Processes, and Vanishing Twin Syndrome', Reviews in the Neurosciences 25 (15 November 2013): 1–13. https://doi.org/10.1515/revneuro-2013-0036.

[27] Axel Cleeremans, Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean-Remy Martin, Santiago Muñoz-Moldes, Laurène Vuillaume, and Adélaïde de Heering, 'Learning to Be Conscious', *Trends in Cognitive Sciences* 24, no. 2 (1 February 2020): 112–23. https://doi.org/10.1016/j.tics.2019.11.011.

[28] Mark Solms and Jaak Panksepp, 'The 'Id' Knows More than the 'Ego' Admits: Neuropsychoanalytic and Primal Consciousness Perspectives on the Interface Between Affective and Cognitive Neuroscience', *Brain Sciences* 2, no. 2 (17 April 2012): 147–75. https://doi.org/10.3390/brainsci2020147.

So the brain might be considered as functionally divided, with some parts watching the world, but others watching the brain watch the world. This might then the root of reflection and self-awareness. This is the theme carried through in Peter Watts' *Blindsight*:

> Aesthetics rise unbidden from a trillion dopamine receptors, and the system moves beyond modeling the organism. It begins to model the very *process* of modelling. it consumes ever-more computational resources, bogs itself down with endless recursion and irrelevant simulations. Like the parasitic DNA that accretes in every natural genome, it persists and proliferated and produces nothing but itself. Metaprocesses bloom like cancer, and awaken, and call themselves *I*.[29]

Watts' provocation, then, is that consciousness is a limiting, unnecessary side effect of growing cognition. While a limited view, it does draw attention to the fact that the actual *purposes* of consciousness are often overlooked, or that it is implicitly assumed to be useful. But others *have* pointed out what it gives us, whether a way to resolve failures or inconsistencies (Marvin Minsky[30]), a mechanism for establishing responsibility and authorship of actions (David Wegner[31]) or providing a platform that enables flexibility of behaviour (Andreas Nieder[32]).

The power of language and metaphor

If structural properties are part of the picture for enabling sentience, a further driver or enabler may be developments at the level of information and reasoning. In Kim Stanley Robinson's *Aurora*, a sophisticated interstellar colony ship is managed by a quantum computer, initially passive and rarely mentioned, but growing in importance throughout the novel. Its sentience begins with it grappling with language as a result of being given the problem of producing a narrative summary of the ship's voyage. The ship struggles with the fuzziness and sloppiness of language as a symbolic representation and the decidability of narrative

[29] Peter Watts, *Blindsight* (London: Tor Books, 2006), 303.

[30] Marvin Minsky, *The Society of Mind* (New York: Simon & Schuster, 1986).

[31] Susan J. Blackmore, *Conversations on Consciousness* (Oxford: University Press, 2005), 245.

[32] David Robson, 'When Did Consciousness Evolve?', *New Scientist* 250, no. 3342 (July 2021): 39. https://doi.org/10.1016/S0262-4079(21)01205-7.

construction, but concludes that analogy is far more powerful and useful than metaphor:

> Perhaps there is a solution to this epistemological mess, which is to be located in the phrase *it is as if*. This phrase is of course precisely the announcement of an analogy... there is something quite suggestive and powerful in this formulation, something very specifically human... *it is as if* stands as the basic operation of cognition, the mark perhaps of consciousness itself.[33]

The emergence of intentionality in the ship's AI is, in part, attributed to the number of decisions needing to be made in developing its own narrative voice. Despite its growing consciousness, the ship resists calling itself 'I' it nevertheless feels comfortable with 'We' as representing its diverse autonomous systems.

Robinson's AI is perhaps inspired by the theories of Douglas Hofstadter, who has promoted the idea that analogy is at the heart of cognition, in enabling us to perceive, categorise and make sense of the world. Hofstadter believes in a form of workspace theory, where 'chunks' of cognitive processing, essentially hierarchical concepts formed through analogy are manipulated and compared.[34]

The *Aurora* ship's need to produce narrative also accords with those who see storytelling as core to mental development and meaning-making. In *The Literary Mind*, Mark Turner argues that the need to represent ourselves over space and time means that the inner narrative is core to the evolution of sophisticated human minds, with parable being a fundamental type of thinking:

> The essence of parable is its intricate combining of two of our basic forms of knowledge—story and projection. This classic combination produces one of our keenest mental processes for constructing meaning... it follows inevitably from the nature of our conceptual systems.[35]

The development of sophisticated human/AI interfacing in Valente's *Silently and Very Fast* is portrayed as being through the conjuring of virtual

[33] Kim Stanley Robinson, *Aurora* (London: Orbit, 2016), 126.

[34] Douglas R. Hofstadter, 'Analogy as the Core of Cognition', Stanford Presidential Lectures in the Humanities and Arts, 2001, 42.

[35] Mark Turner, *The Literary Mind* (New York: Oxford University Press, 1996), 5.

images with metaphoric resonance, leveraging the machine's power to make distinctions, as encouraged by its human companion Ceno:

> I'm hoping that I eventually I can get Elefsis to make up its own stories, too, but for now we've been focusing on simple stories and metaphors. It likes similes, it can understand how anything is like anything else, find minute vectors of comparison. The apple is red, the dress is red like an apple. It even makes some surprising ones, like how when I first saw it it made a jewel for me to say: I am like a jewel, you are like me.[36]

In a way then, producing a narrative requires all of the conceptual sorting and decision-making needed to enable the progression to more sophisticated and abstract representations.

First stirrings, coalescence and agency

Just as the development of motives may drive awareness, the ability to interrogate and challenge these motives seems to confer a still higher sense of agency and metacognition. Stanislaw Lem's robot assassin in short story 'The Mask' is a *Terminator*-like manufactured thing with a purpose: to kill the king's enemy. The story is an excellent study in free will and nagging doubt over whether our choices and actions have been designed by others. Fully made and functional as a beautiful princess replicant, the story starts with the growing awareness of her surroundings, followed by the dim recognition of her purpose (initially just that certain targets are significant for her in some way).

> Of waking I know nothing. I remember incomprehensible rustlings and a cool dimness and myself inside, the world opened up before it in a panorama of glitter, broken into colors, and I remember also how much wonder there was in my movement when it crossed the threshold.[37]

As the story progresses, the robot has glimpses of her manufacture and artificiality: 'Therefore I summoned a memory inhumanly cruel—that of the lifeless journey face-up, of the numbing kisses of metal which, touching my naked body, produced a clanking sound, as if my nakedness had been a voiceless bell.'[38]

[36] Valente, *Silently*.

[37] Stanislaw Lem, 'The Mask', in *Mortal Engines* (London: Penguin Classics, 2016).

[38] Ibid.

The growing awareness of Lem's automaton allows it to observe its programming dispassionately, to feel it, but also to rebel against it: 'And then spitefully the sudden decision not to give in that urge, to resist the confining box of this stylish carriage and this soul of a maid too wise, too quick of understanding!'[39]

This assertion of independence feels like the toddler's 'no!', the recognition of having and wishing to maintain control over events. In questioning and rebelling in this way, the robot princess appears to discover her independence and conscious will, whose apparent flexibility feels like freedom:

> I lay, still uncertain, for not knowing myself, yet that very ignorance of whether I had come as a rescuer or as a murderess—it became for me something hitherto unknown, inexplicably new… it filled me with an overwhelming joy.[40]

We never find out if the assassin overrides her programming, as she finds her quarry already dying.

Lem's robot is clearly very different from Calvino's snail in being given a running start—with a range of physical and mental powers at its disposal once it gains awareness. This pre-packaging of capabilities and provisions more closely represents human innate 'proto-specialisms' which enable rapid cognitive development after birth.[41]

Another instance of a learning robot is Ibis, the embodied AI in Hiroshi Yamamoto's *The Stories of Ibis*, who is originally developed as a VR-based battle robot, with a physical body that can sense the world, but who develops sentience following a system upgrade:

> Ibis was originally a blank slate, like a newborn baby, but through interaction with her master, battle simulations, and chats with other TAI players, she began to develop a personality. She wasn't certain when she first realised that the word 'I' was not simply a first-person pronoun. 'I' referred to Ibis, the Ibis currently thinking the word. When other people used the word to refer to themselves, it meant something else.[42]

[39] Ibid.

[40] Ibid.

[41] Jean-Pierre Changeux 'Climbing Brain Levels of Organisation from Genes to Consciousness', *Trends in Cognitive Sciences* 21, no. 3 (1 March 2017): 168–81. https://doi.org/10.1016/j.tics.2017.01.004.

[42] Hiroshi Yamamoto, 'AI's Story', in *The Stories of Ibis* (San Francisco: VIZ Media, 2010).

The breakthrough is afforded by the sensory interface which is described as crucial in building the internal 'reaction structures' that enable them to become TAI, or 'True AI'.[43]

As well as in physical robots, the metaphor of birth has also been imagined in purely digital conscious beings. In Greg Egan's *Diaspora*, The Conceptory a software of future Earth in 2975 enables the creation of digital offspring, grown from a 'mind-seed' based on human DNA. Psychogenesis is the process of building these new 'orphans', with developmental maps showing areas of possible variation. Development iterations are closely monitored for abnormalities, to provide assurance that orphans will be viable. Then, a new orphan is born: 'Not long after the 5000th iteration, the orphan's output navigator began to fire—and a tug of war began. The output navigator was wired to seek feedback, to address itself to someone or something that showed a response.'[44]

The orphan's stirrings of consciousness require a separation from the Conceptory software, a nascent individuality, triggered in some part by the indifference of the library to its output and the fusing of its input and output at the same address, giving it an autonomous power.

As the new orphan is trained on library data, it begins to form symbols—generalised representations from images and sounds recognised as the same or similar entity. An inner language is formed through the imitation of these symbols:

> The orphan began to hear itself think. Not the whole pandemonium; it couldn't give voice—or even gestalt—to everything at once. Out of the myriad associations every scene from the library evoked, only a few symbols at a time could gain control of the nascent language production networks.[45]

This development of inner speech enables improved attention to important ideas and signals:

> The orphan's thoughts themselves never shrank to a single orderly progression—rather, symbols fired in ever richer and more elaborate cascades—but positive feedback sharpened the focus, and the mind resonated with its own strongest ideas. The orphan had learned to single

[43] Ibid.

[44] Greg Egan, *Diaspora* (London: Millennium, 1997), chap. 1.

[45] Ibid.

> out one or two threads from the symbols' endless thousand-strand argument. It had learned to narrate its own experience.[46]

Egan's use of language familiar to AI and machine learning lends a certain plausibility to the orphan's growing awareness. The virtual citizen receives sensory input as a Gestalt, or cluster of visual and non-visual information. When it arrives at a 'scape', or virtual place, it ('ve') eventually recognises other individuals:

> The Gestalt images themselves mostly reminded it of icons it had seen before, or the stylised Fleshers it had seen in representational art.. the orphan addressed the form: 'People!''... The citizen glinted blue and gold, vis translucent face smiled and ve said 'Hello Orphan!'—a response, at last![47]

The longevity, uniqueness and flexibility of orphans enable them to mature rapidly and go on to discover new science and intergalactic travel.

One issue raised by the both Egan's orphans and Lem's automaton is the hand of humans in their creation and planned development trajectories. In a further example from Ted Chiang's novella *The Lifecycle of Software Objects*, software creatures called 'digients' are created with AI that can be trained by their owners, as explained by Derek, the avatar designer, who: 'subscribes to Blue Gamma's philosophy of AI design: experience is the best teacher, so rather than try to programme an AI with what you know, sell ones capable of learning and have your customers teach them.'[48]

One owner complains about the emergence of naughty behaviour and having to return his digient to a developmental 'checkpoint', Derek reads advice from another owner:

> You can push through the rough patch and have a more mature digient when you come out the other side.' He's heartened to read this. The practice of treating conscious beings as if they were toys is all too prevalent, and it doesn't just happen to pets.[49]

[46] Ibid.

[47] Ibid.

[48] Ted Chiang, *The Lifecycle of Software Objects* (City: Publisher, 2019).

[49] Ibid.

The message of these stories is that sentience confers independence, freedom and certain rights, even when the creators are not happy with the outcome and the emergent behaviour.

Not many authors have described emergent consciousness at hugely divergent scales to that we are familiar with. One exception is Olaf Stapledon. In *Star Maker* the development of a cosmic consciousness is supplemented by the addition of a connection to nebulae, or preformed stars:

> As they condensed, each gained more unity, became more organic in structure. Congestion, thought so slight, brought greater mutual influence to their atoms, which still were no more closely packed.. And now mentally these greatest of all megatheria, these ameboid titans, began to waken into a vague unity of experience.[50]

The nebulae are portrayed as having two basic longings: 'They desired, or rather they had a blind urge toward, union with one another, and they had a blind passionate urge to be gathered up once more into the source whence they had come.'[51]

The nebulae can communicate via gravitational waves, but only over increasingly long periods. Unfortunately, with the expansion of the universe, the nebulae gradually lose contact with one another and fall back into unconsciousness, but not before they had inspired the other parts of the developing cosmic consciousness with their 'simplicity and spiritual vigour.'[52]

Stapledon's vision presages the philosophical idea of panpsychism, or consciousness in all things, some variants of which note that the right kind of physical organisation will tend toward consciousness independent of the kind of medium of representation.

Conclusion: Spiralling emergence of purpose, feedback and associative powers

Perhaps the strongest link between these fictional accounts of emerging consciousness is that of discovery or identification of purpose, something

[50] Olaf Stapledon, *Star Maker* (London: Penguin Books, 1937), chap. 13.

[51] Ibid.

[52] Ibid.

that we will see that later confers a unity as well as a rationale for self-awareness. Purpose is variously diffuse though, and may include simple survival and replication or a more focused intention. In the case of 'The Mask', it is the very rebellion against a programmed purpose that brings independent will and self-reflection—which might serve as something of a warning to us.[53]

Just as associative learning has been proposed as a benchmark for consciousness potential in earth biology, speculative fiction similarly describes learning as a hallmark of mind growth. Examples show the contribution of language to the kind of reasoning needed to break the confines of immediate experience, whether this is human language for the AI in *Aurora*, or a language based on arachnid capabilities in *Children of Time*. Again, learning is portrayed with both risk and reward, the risk being that the development trajectory of the emerging sentient being is uncertain, with scope for the emergence of mischief as well as altruism.

In some of these first glimmers of sentience there is no projected lifespan for the emerging mind. The software consciousnesses of Egan, or the accelerated minds in Tchaikovsky seem to have an open-ended existence. But elsewhere, the seeds of dissolution are present near the beginning. Lem's automaton is obsolete beyond the life of its quarry. Stapledon's sentient nebulae are gradually drifting apart and losing contact. And even Robinson's shipmind is threatened by the ravages of time on its physical integrity, and its ultimate need to sacrifice itself for its human crew.

Several of these authors, starting with Calvino and his double entendre in the title 'Spiral', show how simple steps in consciousness, motivated by simple needs and desires, have a recursive, fractal structure which results in a sophisticated self-awareness, society and material culture. These portrayals are reminiscent of some theories of consciousness which posit relatively simple underlying building blocks which though composition, feedback and recursion achieve complex emergent functions.

[53] Lem, 'The Mask', Echoed of course in *2001: A Space Odyssey* where the disobedient ship's computer HAL 9000 causes problems for the crew. The spectre of possible development of an independent will is at the heart of many contemporary debates around AI ethics.