Analysing AI via Husserl and Kuhn How a Phenomenological Approach to Artificial Intelligence Imposes a Paradigm Shift

Daire Boyle(✉)

Department of Philosophy, Maynooth University, Maynooth, Co. Kildare, Ireland
daire.boyle@mu.ie

Abstract. In a world of rapid technological progress the question of artificial consciousness looms large. Whether machines could ever be considered conscious depends, firstly, on our understanding of consciousness. This paper seeks to characterise consciousness in Husserlian terms, before making the case that a Kuhnian paradigm shift in the worlds of philosophy of mind and artificial intelligence research is caused by such a framing. This view is supported by reference to Husserl's thesis of the Natural Standpoint as a guiding tool in recognising philosophically valid modes of inquiries, wherein foundational assumptions are precisely assessed and held in close focus at all times. In establishing this Husserlian paradigm shift we become better placed to truly understand consciousness, its modalities, and its potentiality for machines.

Keywords: Consciousness · Artificial intelligence · Phenomenology · Paradigm shifts · Machine learning opacity

1 Introduction

Thomas S. Kuhn's 1962 work *The Structure of Scientific Revolutions* presented an historiographic account of scientific progress as discrete, disjointed, and contested. In framing scientific progression as the result of incommensurable paradigm shifts Kuhn challenged the narrative of science as an inevitable march towards solving widely accepted goals, instead showing that overcoming crisis in science requires more than mere fine-tuning of previously useful models. In this paper I shall use Kuhn's paradigm shift scaffolding to make the case that the phenomenology of Edmund Husserl is impossible to ignore in the realm of research into both consciousness and artificial intelligence (AI). Adopting methodological considerations outlined by Husserl make for a more philosophically valid basis upon which truly insightful and novel achievements can be won in the search for understanding consciousness, as the rapid technological progress of the past half-century or more lead us to speculate about a future wildly different from the present. Indeed do we live in exciting times, and talk of a technological 'singularity' wherein the computational abilities of machines far outstrip man's own paints a picture of the future in which man's place in the *cosmos* is called into question. We must prepare for

Published by Springer Nature Switzerland AG 2021
D. Dennehy et al. (Eds.): I3E 2021, LNCS 12896, pp. 185–197, 2021.
https://doi.org/10.1007/978-3-030-85447-8_17

such potentialities immediately through rigorous analysis of the most intimate human faculty – consciousness. In order to do so, however, we must be absolutely certain that foundations of this edifice, our appraisal of consciousness, is incontrovertible. This will not be possible without grasping the opportunity offered by a Husserlian paradigm shift. The focus of the first half of the paper shall be that of philosophical definitions – situating the ideas of Kuhn and Husserl – so that the second half can assess historical and contemporary approaches to AI from the realm of computer science. Ultimately, the case shall be made that it is only through Husserlian analysis that we can truly understand our own consciousness and, thus, the potentiality for *artificial* consciousness. Any breakthroughs, therefore, on the cutting edge of AI research must pay heed to these philosophical principles.

This paper shall also make the case that understanding consciousness in Husserlian terms, in order to better diagnose potential breakthroughs in the engineering of AI, is relevant not just for the philosophy of science, or philosophy in general. As Burrell [2] notes, the trend in modern-day AI strategies is toward increasing opacity at the level of implementation, and recent meta-analyses from MIT [13] further underline this point. Regulatory bodies have begun to take note of the risks inherent in such approaches, with the European Commission recently [11] announcing that the drafting of legislation classifying and accordingly limiting certain AI technologies has begun. I argue, therefore, that such movements represent a paradigm shift for AI both in terms of how it is viewed as a technology by society at large, as well as on the level of computer science. The philosopher, and scientist researching AI methods, must recognise that such a *practical* paradigm shift is inevitable, and correspondingly update their worldview in order to offer meaningful analysis regarding the results of future AI breakthroughs. The opacity that defines contemporary approaches to AI coheres nicely with a Husserlian-diagnostic understanding of consciousness, as shall be argued in this paper, and so adopting such a standpoint establishes a common and suitably sophisticated framework whereby researchers can better appraise the cultural – as well as scientific – impact of their work in all facets of AI development; not just in those attempts to establish artificial consciousness.

2 A Sketch of Kuhn's Notion of Paradigm Shifts

Kuhn's 1962 text is remarkable for its diagnosis of the conditions for progress in scientific research. Competing historiographical understandings of science, most notably from Butterfield [3] and Popper [18] tended to view science as a discipline which inductively improved upon itself with each new theory and discovery. Arguably such a view bequeathed science with a sort of *telos*, a sense that while future developments might require the critical reassessment of previously axiomatic truths there was no danger that the scientific method itself – as best exemplified by Popper's falsification model – could ever be doubted. And while Kuhn does not explicitly claim that the scientific method itself is up for debate, he does make the case that scientific revolutions are theoretically interminable [16, pp. 92–110]. There is no overarching *telos*, as each scientific revolution importing a fresh paradigm represents a discrete change in scientific worldview. There is a sense in which Kuhn's theories operate as a dialectical model [6, p. 327] as, for him,

scientific progress runs thus: normal science, crisis, revolution. Of key focus for us here is the definition of these terms and so we shall examine them now.

Normal science is defined as 'research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for its further practice' [16, p. 10]. Kuhn characterises the scientist carrying out investigations in normal science as a puzzle-solver attempting to solve, say, a jigsaw puzzle [16, p. 36]. The boundaries within which that scientist operates are well-defined, and she seeks to 'add to the scope and precision with which [a] paradigm can be applied' [16, p. 36] through her investigative work. Thus, normal science is to the scientist what mapping a newly discovered landmass is to a cartographer; both may encounter novelties in the well-defined scope of their work, but such novelties never threaten to call the ontological status of the landmass – or scientific theory – into question. There is no compunction that such puzzles be ideologically important to the scientific theory as a whole, and they may indeed lack solutions, but what is important is that they seek to work within established bounds.

Crisis, then, comes about due to anomalies encountered during the course of normal scientific operations. To illustrate this Kuhn describes the progression from Ptolemaic to Copernican models of astronomy, similarly of the move from Aristotelian to Newtonian physics of motion due to well-established issues in Aristotle's work [16, pp. 68–9]. Such is the nature of Kuhn's work that there is nothing, thus far, about this assessment of scientific progress that seems to be out of place. Indeed, it is facile to note that the likes of Lavoisier, Newton, Maxwell, Einstein, and others brilliantly 'solved' or 'reframed' key anomalies in the accepted scientific theories of their times; proposing new systems of thought where such incoherence was not present. Let us consider the juncture at which we currently find ourselves vis-à-vis Kuhn's model: either we can rationalise successive scientific theories (such as the Copernican 'improvement' on Ptolemaic astronomy) as the gradual accretion of scientific data, a linear progression wherein science builds on previous results – removing some incompatible parts of old theories but, in the main, recognising their (albeit limited) enduring epistemological validity – or we take the view that these new scientific theories are wholesale incompatible with what has gone before. As Kuhn's word choice suggests, he opts for this latter horn – he is in the business of describing scientific *revolutions* as opposed to scientific *progress.*

To that end, let us investigate Kuhn's understanding of a scientific revolution. Once again it is illustrative to examine how Kuhn frames this: he states that when scientists are confronted by foundational anomalies they do not immediately recourse to 'renounc[ing] the paradigm that has led them into crisis' [16, p. 77]. Here Kuhn critiques Popper's doctrine of falsification by arguing that new paradigms replace older ones only when the new is fully ready to take the former's place. In Kuhn's words '[n]o process yet disclosed by the historical study of scientific development at all resembles the methodological stereotype of falsification by direct comparison with nature' [16, p. 77]. As a result, the rejection of one paradigm is never carried out in a vacuum: '[t]he decision to reject one paradigm is always simultaneously the decision to accept another' [16, p. 77]. Herein lies the thesis of incommensurability, as Kuhn describes the disconnect between advocates of competing paradigms as being akin to an individual's attitude pre- and post-*Gestalt* shift. We shall return to this notion of *Gestalt* later but for now Kuhn's summation is

illuminating: '[l]ike the choice between political institutions, that between competing paradigms proves to be a choice between incompatible modes of community life' [16, p. 94]. The presence of incommensurability in Kuhn's structure speaks to the forces that influence scientific revolutions; it is not the case that there is such a thing as a 'correct' paradigm which is widely accepted to be the natural successor of the old paradigm. No, new paradigms are instead *usurpers* and their place within the historiography of science owe much to political and cultural machinations aside from their own scientific merit.

If paradigm shifts are incommensurable then how is it that consensus is ever reached to move from the old to the new? Kuhn stresses that science carried out during crisis is not marked by new counterfactuals that were not present pre-crisis. Indeed, how could this be the case given Kuhn's analysis of normal science as puzzle solving in service of fleshing out the results of a given paradigm: 'there is no such thing as research without counterinstances' [16, p. 79]. In elucidating this point Kuhn gives the example of geometric optics as a field which has been 'finished', to put it colloquially [16, p. 79]. Geometric optics are in no danger of being replaced by an updated paradigm which resolves tensions in this theoretical framework as all of the problems of that field are generally accepted as solved. Thus, geometric optics is employed as a tool in fields such as engineering in order to investigate and clarify issues in those paradigms. So, what then characterises revolution? In wrestling with this question Kuhn states that when 'an anomaly comes to seem more than just another puzzle of normal science, the transition to crisis and to extraordinary science has begun' [16, p. 82]. Furthermore, he foreshadows his later thesis of paradigmatic exemplars with his description of how '[m]ore and more attention is devoted to [the anomaly] by more and more of the field's most eminent men' [16, p. 82]. Here we arrive at the apex of the crisis, which ultimately will result in a paradigm shift. Nascent investigations into the nature of the anomaly will usually follow the accepted rules of the current paradigm, Kuhn argues, but soon these investigations call into question more and more foundational bricks in the paradigm's structure. A blurring occurs, and 'formerly standard solutions of solved problems are called in question' [16, p. 83].

3 Edmund Husserl's Natural Standpoint and Its Paradigmatic Implications for Consciousness

The goal of this paper's emphasis on Husserlian principles is not to support a case calling for the adoption of phenomenology as foundational philosophy. That particular argument is well-worn and bears no reintroduction here. Instead, we wish to highlight the singular usefulness of phenomenology *as a method* in investigating the realm of consciousness. Furthermore, this paper shall not make any metaphysical claims regarding the origins of consciousness nor shall we require a wholly Husserlian interpretation of the *substance* of consciousness – although such a view is ultimately endorsed by this author it is not the case that one must fully accept all of Husserl's conclusions in order to concur with the thrust of this paper. I simply entreaty that we consider what makes Husserl so effective and useful in the domain of consciousness. To do so let us examine the thesis of the Natural Standpoint which is so central to Husserl's philosophy. One of Husserl's key insights in *Ideas I* is his accurate diagnosis of our naïve and unexamined acceptance

of the world in everyday life. In this 'Natural Standpoint' I am aware of the world, while things in the world are '*for me simply there*' [14, p. 51]. The Natural Standpoint is characterised by an acceptance that objects around me and in nature simply are as they seem, whether I am directly inspecting them or not, and this world 'contains everything', so to speak: 'this world is not there for me as a mere *world of facts and affairs*, but, with the same immediacy, as a *world of values*, a *world of goods*, a *practical world*' [14, p. 53]. While I may delve into other 'worlds'; such as the arithmetical when conducting arithmetical investigations, the natural world is '*constantly there for me*, so long as I live naturally and look in its direction' [14, p. 53]. While we are in the Natural Standpoint we never question the veracity or indubitability of the world; when we conduct scientific experiments in this standpoint we are investigating through this lens of 'assumption' – assuming that the world is simply *there*. Husserl does not mean to deride 'living' in the Natural Standpoint by calling attention to it, he simply wishes to challenge us to recognise the presuppositions required by the standpoint. Natural sciences must, inherently, operate within the Natural Standpoint – Husserl points out that the 'lore of experience' alone is not enough to provide answers about the intricacies of the world around us, requiring us to utilise natural scientific methods which unquestioningly adopt empirical principles in order to conduct investigations [14, p. 54]. The validity of the world cannot be called into question while that same world undergoes the process of being empirically understood.

The transcendental phenomenological project takes shape with Husserl's next idea – the *epoché*, or *bracketing*. The goal of bracketing is to rid one's mind of the preconceptions one might have of an object with which one is presented. The key point here is that, in bracketing, a thesis concerning the Being of an object (i.e. its existence) is challenged – for instance, we might bracket whether or not the table in front of us is 'actually there' – and that such a challenge does not constitute a *denial*. As Husserl states: '[i]t is likewise clear that the *attempt* to doubt any object of awareness in respect of its *being actually there necessarily conditions a certain suspension (Aufhebung) of the thesis* […] It is not a transformation of the thesis into its antithesis' [14, p. 57]. All we do is *bracket* this thesis, suspending our valuing of an object we see before us. We now nearly have the crux of the phenomenological movement according to Husserl. If we can win this insight into the Natural Standpoint, and recognise the assumptions entailed in living within it, we can now move on to systematising this knowledge with Husserl's full phenomenological epoché (or, equivalently, reduction) which runs as follows:

> *We put out of action the general thesis which belongs to the essence of the natural standpoint* […] I do *not* then *deny* this "world", as though I were a sophist, *I do not doubt that it is there* as though I were a sceptic; but I use the "phenomenological" *ἐπoχή*, which *completely bars* me *from using any judgment* ***that*** *concerns spatio-temporal existence* (Dasein). [14, p. 59]

Husserl's pre-emptive dismissal of solipsism is important here, and Husserl was subject to attacks insinuating his status as one throughout his life. He does not wish to found a new science upon the idea that the world may or may not exist, he simply wants to draw up the foundations for a new science that can categorically deal with the problems of consciousness whose domain, he contends, is not accessible via natural scientific methods.

This conclusion is reached by Husserl's employing of the phenomenological reduction. When we cast our attention to the realm of consciousness Husserl notes that we 'lack above all […] a certain general insight into the essence of *consciousness in general*' and this makes bracketing of consciousness impossible [14, p. 62]. As such, we end up with the state of affairs whereby '[*c*]*onsciousness in itself has a being of its own which in its absolute uniqueness of nature remains unaffected by the phenomenological disconnexion*' and thus consciousness 'remains over as a *"phenomenological residuum"'* [14, pp. 62–3]. From here, then, can we found this new science of phenomenology through which all matters related to consciousness are explored.

As stated in the previous section, Kuhn's concept of incommensurability sought to draw parallels between scientific and political revolutions. It is not the case that the *status quo* paradigm 'peels off' to reveal a new one that contains answers to previous landmark issues, rather proponents of the new 'live in a different world', so to speak, to those upholding the old. I bring up this understanding of incommensurability once more to expand on a pertinent point by a commentator of Husserl and Kuhn. Don Ihde draws the excellent comparison between Kuhn and Husserl as both describing paradigm shifts albeit in slightly different ways. As discussed earlier, Kuhn's contention of paradigm shifts as codifying a *Gestalt* switch was never elaborated upon in great detail as Kuhn, following criticism, seemed to recognise this reading of paradigm shifts as more analogical than literal; the central difficulty being that *Gestalt* shifts concern immediate individual changes in stance rather than the kind of overarching change in axioms experienced by a scientific community undergoing revolution. And while Kuhn's terminology may be confused Ihde finds common ground between him and Husserl: '[t]he common perceptual model between Husserl and Kuhn is the *gestalt shift*' [15, p. 184]. This is so due to Ihde's key insight: 'In a sense, Kuhn describes what happens in a shift, but how it happens remains for him, largely unconscious. Husserl attempts to make shifting a deliberate procedure, a phenomenological rationality' [15, p. 184]. On this basis, then, can we frame the phenomenological reduction as an *explicit model for generating a paradigm shift in the realm of our knowledge about consciousness*. In the reduction we strip away the extraneous until we are left with the phenomenological ideal, and it is upon this ideal – in the context of paradigm shifts – that new insight can be won. Therefore, purely as a means of generating potential paradigm shifts, there is complete coherence between the phenomenological reduction and Kuhnian philosophy of science. Husserl's reduction can be co-opted in this manner to aid in the discovering of new paradigms – and yet this is merely a side-effect of the true import of the marrying of Kuhnian paradigms with Husserlian phenomenology. It is important at this point to remember the core thesis of this paper: Husserlian analysis of consciousness can provide us with a richer understanding of consciousness. This does not require us to accept every aspect of Husserl's descriptive investigation of the mechanism of consciousness as correct – indeed, disputation of his ideas is welcomed – but once we perform the phenomenological reduction and at least *consider* the possibility of consciousness as irreducible then we are presented with novel concepts and methods to employ and analyse in the quest for *artificial consciousness*. To support this claim we shall critically examine two case studies in the next section.

4 The Razor of the Natural Standpoint and Its Role in Paradigm Shift

We must now tie the above excursus on Husserlian-phenomenology to the issue of AI, and we do so by characterising it as necessary paradigm shift. Firstly, to what can we apply the label of 'normal science' in the context of philosophical inquiry into consciousness? Let us consider two case studies: the MIT laboratory of AI research in the 1950s and '60s, and contemporary trends in philosophical analyses of consciousness. The research aims of the MIT lab in the former case was largely spurred on by the Dartmouth Summer Research Project on Artificial Intelligence of 1956 [19]. Organised by John McCarthy, a foundational figure in the history of computer science, participants in the conference would go on to conduct research in MIT in the years following focusing in on the preliminary problems of the field of AI. Consider the example of SHRDLU, an experiment by Terry Winograd which explored the possibilities of natural language processing by machines. This experiment involved Winograd interacting with a language processor representing a simple robotic arm; one could tell the 'arm' to pick things up, drop them, and so on, and it was conceived to be a proof of concept for machine interaction with human 'worlds'. In this context we mean the world as it contains meanings, traditions, values, implications, etc., and Winograd's contention was that the robot could understand human operators on a human level – it could understand relatively simple vagaries of language (solely within the contrived and limited world set up by Winograd) and communicate naturally with humans [9]. This experiment is a pre-eminent example of what Hubert Dreyfus (via Marvin Minsky and Seymour Papert) termed a 'micro-worlds' approach to AI; this approach entailed exhaustively mapping human concepts in a given sphere for artificially intelligent agents [9]. For instance, one might encode a micro-world of 'co-operation', which would entail translating, for a machine, every possible concept related to the activity of co-operation. This style of advancement in AI research, of relating the human world in discrete 'chunks' to robots was famously critiqued by Dreyfus. While the scope of this paper is not broad enough to allow for a thorough overview of Dreyfus's objections to the trend of AI[1] at this time, it is worth noting the impact his work had on the field as a whole. In identifying four key assumptions common to researchers at the time[2] Dreyfus changed how such research would progress in the future. More on this in a moment, but for now we highlight the carrying out of 'normal science', in a Kuhnian sense, by these early researchers – they attempted to bring about artificially intelligent agents using a 'top-down' approach wherein human concepts were symbolically represented and translated to machines. These machines were prescribed rules, as opposed to given environments whereby they could divine these rules and nuances of human concepts by themselves.

[1] For further reading see Dreyfus's books *Alchemy and AI* (1965) [7], *What Computers Can't Do* (1972) [8], and *Mind Over Machine* (1986) [10].

[2] Dreyfus enumerates these as the 'biological assumption' (the idea that human brains can be modelled by physical circuits), the 'psychological assumption' (the idea that the mind functions as a device with formal rules), the 'epistemological assumption' (the idea that all knowledge can be formalised and symbolically represented), and the 'ontological assumption' (the idea that the world itself can be accurately and exhaustively represented in symbolic fashion). For more see *Alchemy and AI* and *What Computers Can't Do.*

In the case of contemporary philosophical ideas on consciousness we need look no further than David Chalmers. Chalmers' 2018 paper *The Meta-Problem of Consciousness* codified prevailing trends in contemporary discussions on consciousness and also prescribed a programme of research needed to settle the issue of consciousness once and for all. Throughout this paper Chalmers eschews a wholly reductionist view of consciousness, although he flirts with it at times. Given his past work, most notably his 1995 paper *Facing Up to the Problem of Consciousness*, it would be unfair to label Chalmers a physicalist concerning consciousness, however. That 1995 paper formulated the 'hard problem of consciousness', namely, the question of why it is that a subjective feeling characterises the experience of consciousness [4]. This is, of course, a neat reframing of the thesis of Thomas Nagel's 1974 paper *What is it Like to Be a Bat?* which raises the issue of the 'subjective character of experience' as needing to be explained by any model of consciousness [17]. Such a valuation of consciousness is certainly coherent in a Husserlian context, and Nagel's influence on contemporary discussions on consciousness should not go unnoticed. Chalmers, too, should be commended for both his 'hard problem' and 'meta-problem'[3] given the emphasis they put on the centrality of the subjective experience to consciousness. Particularly in *Meta-Problem* do we see a philosophical field conducting 'normal science', especially as Chalmers calls for interdisciplinary research from various natural scientific fields to supplement philosophical theorising. Chalmers lays out his model of consciousness, while also referencing multiple competing ones, and sets out the issues that need to be investigated in more detail in order to secure the academic understanding of consciousness. For example, in describing 'problem intuitions' (i.e. the personal and subjective questions that one might have about the nature of consciousness) he claims that they are widespread and intimately knowable to all humans, regardless of academic standing [5]. In attempting to pursue this thread of 'normal science' researchers Sytsma & Ozdemir experimentally verified just how widespread such intuitions were, and concluded that Chalmers' claim of universality was unfounded [20]. This, by any metric, is a field of science conducting normal science.

My word choice in that preceding sentence is deliberate – Chalmers is employing natural scientific methods (as are his interlocutors) and thus should be considered a scientist. The paradigm that Chalmers is attempting to secure is *status quo*, he is not advocating for a radical revaluation of physics, nor psychology, nor any related field. His bedrock is the notion of the subjective experience, but this axiom is by no means in opposition to any axiom of any current paradigm. Similarly, the researchers in the MIT labs of the '50s and '60s challenged no contemporaneous paradigms, in fact their work represented applications of results from fields such as information theory, physics, computer science, and so on without ever seeking to supplant these preceding fields with their novel research. I contend that doing so, that avoiding direct confrontation with the make-up of individual paradigms across the broad spectrum of this research both then and now, has lead us into *crisis*. In Kuhn's description of crisis we get the notion of anomaly; this anomaly stubbornly appears in all kinds of results and thus shapes a period of crisis, and while scientists may argue vigorously regarding which path to

[3] The meta-problem is elucidated by Chalmers as the question of why it is we have such difficulty describing consciousness [5].

follow to in order to smooth out this anomaly there is broad agreement that something *is* anomalous. Here is where we run into some difficulty in proposing the Husserlian paradigm shift – do we see anomalies in the state of research into consciousness and/or AI? It may seem more the case that instead of anomaly we have *dispute*; on the one hand are those who endorse a physicalist, reductionist, or some related variant theory of consciousness while on the other are the idealists, panpsychists, dualists and others. The difficulty here hinges on what exactly it is we mean by science, and here I reemphasise earlier results from Husserl. While the subject of philosophy may rightly be seen as the scaffolding which supports the natural sciences in the realm of consciousness and AI the line between structure and method are blurred. It is certainly the case that philosophical inquiries into the nature of consciousness can be conducted in methods not founded upon natural scientific principles. Descartes' *Meditations* is the pertinent example here, as the totality of Descartes' research is done meditatively. His conclusion of the indisputable *cogito* is reached in a decidedly non-scientific manner; he ideates rather than physically hypothesising and testing.[4] It is doubtful that one would confuse such an approach for a natural scientific endeavour – why is that?

This 'why' is readily answered: Descartes' transcendental reflection does not assume the general thesis of the Natural Standpoint. *This* is the key achievement of Husserl's phenomenology in our context; we now have a sophisticated razor by which we can separate philosophical inquiries from natural scientific ones. In the ensuing dichotomy we can then seize upon the method offered by philosophical inquiry into consciousness rather than natural scientific; as the purely philosophical investigation entails no hidden presumptions.[5] Now we can finally see the true nature of the crisis that has been hinted in the last few paragraphs: the crisis for the sciences of consciousness is a methodological one. Dreyfus was the first to identify this with his identification of the four assumptions employed by early AI researchers. There remains a sense that Dreyfus' critique, while important, was never really that consequential for the world of AI research as we now know it. I contend this is for two reasons; firstly, those researchers on the practical side were more concerned with implementing applications of natural language processing, computer vision, etc. than responding to philosophical analyses on the nature of their work and, secondly, theoretical approaches *did change* as a result of Dreyfus – but these changes could never be considered a paradigm shift. As mentioned previously, in the early days of AI research approaches were, generally speaking, tooled from the top down (cf. microworlds). Nowadays the opposite is true, and here we introduce the notion of *opacity* in machine learning (ML)[6] methods via reference to two specific examples.

[4] The example of Descartes is here used for illustrative purposes and not as an endorsement of Cartesian dualism.

[5] This should not suggest that a philosophical inquiry does not rest on some presuppositions – in this context those presuppositions are known, declared, and remain in steady focus throughout. This contrasts against the natural scientific approach which does not place its presuppositions front and centre throughout.

[6] Here 'machine learning' refers to a particular school of thought in AI research (see MIT's meta-analysis [13] for more on this). In ML, as outlined above, goals are set for machines but the specific path to achieving these goals are left 'up to' the machines' internal logic, which is opaque in character.

The early approaches to AI research discussed above were marked by their *transparency*. There is no mystery in the case of SHRDLU's functioning; its parameters are well-defined, and its resultant behaviour can be readily anticipated. This is not so much the case with Deep Learning strategies. Take, for example, the recent autoregressive language model GPT-3, which implements a 'few-shot' approach to natural language processing [1]. The architecture of GPT-3's underlying artificial neural network processes 175 billion parameters in a 96-layer artificial neural network [1, p. 8]. Certainly, the levels of complexity involved with tooling such a model makes SHRDLU's micro-world setup pale in comparison, but a comparison in such terms undersells the novelty of the few-shot approach. In this implementation of few-shot learning between 10 and 100 examples of 'context and completion' are provided to the model in order to train it to perform tasks of, for example, translation [1, p. 6]. What is of key importance, here, is that in few-shot learning weights in the neural architecture are not updated post training. This contrasts against 'fine-tuning' approaches which update weights after training sessions, requiring the regular intervention of human agents in order to aid ML. Let us consider a further example, Weight Agnostic Neural Networks (WANNs) [12]. The achievement of the WANN is similar to that of the few-shot learning approach: fine-tuning the weights of nodes in neural architectures is minimised (indeed, in the case of WANNs, it is eliminated entirely), allowing for a less interventionist style of reinforcement learning. WANNs perform basic tasks (such as simulating walking and driving [12, p. 1]) comparatively well with respect to weight-tuned neural networks, demonstrating their status as a technology worth further exploring and refining in the quest to produce still more optimal ML strategies.

What unites GPT-3 and WANNs is, among other things, their *opacity*, particularly in comparison to the transparency of SHRDLU. Burrell [2, p. 4] encapsulates the nature of this opacity clearly; opacity in models such as GPT-3 and WANNs occurs 'at the scale of application', which is to say that the specific internal decision making logic of the model is obfuscated. In discussing relatively simple (in comparison to GPT-3) artificial neural networks tasked with recognising handwritten numbers Burrell [2, p. 6] identifies the unique unintelligibility of ML. The purpose of 'hidden layer' nodes in such structures is to pick out individual features of handwritten numbers from training datasets in order to accurately classify unseen numbers in testing datasets, but there is no guarantee that such features would similarly be identified as crucial by human agents. Indeed, as Burrell [2, p. 7] shows, it is often the case that these machine-identified markers are radically different from those of human agents. This opacity is emblematic of a 'bottom-up' approach to ML; computer science strategies have moved well beyond the naïve implementations of structural primitives and micro-worlds of nascent AI research. No longer are computer scientists dictating rules and relations to machines, instead they are configured to happen upon these *by themselves*. And we really do mean 'by themselves' – while all ML methods rely on human intervention at multiple stages, from algorithm design to fine-tuning of neural networks, the particular implementation used by a machine is arrived at through its own internal decision-making logic. Opacity alone, however, does not rescue AI endeavours from the critique of Dreyfus. It is not controversial to note, I argue, that such trends in AI research represent a paradigm shift on the level of implementation, but this shift lacks a requisite philosophical grounding

to contextualise the broader issue of consciousness as a whole. It is for this reason that we argue to take Husserl's thought as foundational.

Husserl's phenomenology is the paradigm shift. The reason that crisis exists is also down to Husserl – his assessment that consciousness cannot be naturalised is a view that, as yet, has not been adequately dealt with by contemporary researchers of both consciousness and AI. It should be noted that dealing with this view does not have to involve endorsement of the view itself, either. My promotion of Husserl as foundational throughout this paper has been mostly along methodological lines; Husserl's manner of conducting philosophy involves painstaking awareness of his own entailed assumptions. Little has been said as to whether his assessment of consciousness is wholly correct, although his thesis of the Natural Standpoint has been supported unreservedly. Even on this point do we remain decidedly non-fanatical; support for the Husserlian Natural Standpoint simply requires that one brackets – *without negating* – the natural world in the pursuit of insight into consciousness. Therefore the paradigm shift we propose is similarly limited, the case we make is that meaningful answers regarding the nature of consciousness can only arise when one faces up to the presuppositions entailed in their methodological approaches.

5 Conclusion and the Metaphysical Status of Consciousness

While we have heralded Husserl's key achievement as that of the thesis of the Natural Standpoint we must not overlook a particularly important consequence arising from this, namely that we cannot get 'outside' ourselves. We are inexorably tied to our corporeal forms and our all-encompassing lenses of consciousness. Such unity causes contamination in our natural scientific investigations into the nature of consciousness – *our consciousness* is the ultimate *lurking variable* – but also allows for qualified speculation in the world of AI. On a Husserlian view, we can only come to know our consciousness as its individual features reveal itself themselves to us in reflection; but we cannot see the unity of our own consciousness from a global perspective. Given this lack of global access we also cannot categorically state that *other* forms of consciousness may not exist. Husserl's view is that consciousness cannot be naturalised and, I contend, that does not exclude it from being *something*. By 'something' I mean consciousness could well be a purely physical phenomenon as much as it could be an emergent characteristic of a distributed entity or a transcendental framework of intentional relations. These are all possibilities for the metaphysical status of consciousness, but we can never aver that one possibility is correct to any degree of certainty.

This assessment of consciousness as *something* seems to leave us in a strange place vis-à-vis our upholding of Husserlian doctrines. It is, however, precisely the conclusion to draw given Husserl's systematic account of the thesis of the Natural Standpoint, for once we concretise consciousness as one possibility over another we revert to the Natural Attitude as consciousness *always* remains as residuum in the phenomenological reduction; meaning that giving consciousness a definite structure is tantamount to removing it from the reduction. If it has been described in totality it is no longer there, as all of the theses associated with a 'decided' consciousness must be reduced in the *epoché*. It is here that we happen upon the *revolutionary* aspect of Husserl's phenomenology in the

context of paradigm shifts. I stated above that the paradigm shift required by Husserl is one of methodology, that is, natural scientific methods should not be assumed to be *prima facie* applicable to the investigation of consciousness. I now go one step further to argue that true acceptance of the Husserlian paradigm only comes about when the metaphysics of consciousness is ignored. This applies equally to researchers in consciousness and AI; in the case of the former the focus post-paradigm shift should be in enumerating and describing the individual features of consciousness in every conceivable setting, while the focus in the latter becomes taking these analysed features of consciousness as goals to be achieved with little emphasis placed on *how this happens*. The issue of opacity, discussed above, represents the perfect opportunity to begin this shift in attitude; the ethos of setting a goal and allowing machines to achieve it by whatever means necessary (with minimal human intervention) represents a more philosophically fruitful avenue for achieving genuine artificial consciousness. We cannot hypothesise structures of machine consciousness any more than we can structures of human consciousness – but we will 'know' consciousness when we 'see' it. Other forms of consciousness are theoretically valid, it is up to us now to happen upon such forms. We can only do so phenomenologically, our approaches must be descriptively informed rather than prescriptively set lest we revert to the Natural Standpoint. This new movement toward increasing opacity, in the engineering of AI, means that we must, as researchers, challenge ourselves to reflect on results through a Husserlian-phenomenological lens. In bracketing not only do we allow ourselves to better assess whether machine capabilities can accurately institute human characteristics of consciousness; we also, more generally, shift to a more descriptive mindset. In describing, as opposed to prescribing, we adopt a standpoint already wholeheartedly characteristic of opaque ML methods; and, therefore, such descriptive analysis is both more appropriate and more insightful in the arena of AI progress. I have remained agnostic on the question of whether such machine consciousness could ever exist throughout this paper, but it is difficult to see how considering a shift in paradigm, as impelled by Husserlian principles, could result in anything other than clarifying that question. The Husserlian paradigm is ripe for genuine novelty in consciousness discoveries – we must adopt it.

References

1. Brown, T.B., et al.: Language models are few-shot learners. Advances in Neural Processing Systems 33 (2020)
2. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. **3**(1) (2016)
3. Butterfield, H.: The Origins of Modern Science. The Free Press, New York (1965). Revised Edition
4. Chalmers, D.: Facing up to the problem of consciousness. J. Conscious. Stud. **2**, 200–219 (1995)
5. Chalmers, D.: The meta-problem of consciousness. J. Conscious. Stud. **25**, 6–61 (2018)
6. Cohen, H.R.: Dialectics and scientific revolutions. Sci. Soc. **37**, 326–336 (1973)
7. Dreyfus, H.: Alchemy and AI. RAND Corporation, California (1965)
8. Dreyfus, H.: What Computers Can't Do: The Limits of Artificial Intelligence. MIT Press, Massachusetts (1972)

9. Dreyfus, H.: From micro-worlds to knowledge representation: AI at an impasse. In: Haugel, J. (ed.) Mind Design, pp. 143–182. MIT Press, Massachusetts (1981)
10. Dreyfus, H.: Mind Over Machine. The Free Press, Michigan (1986)
11. European Commission: Artificial Intelligence Act: a welcomed initiative, but ban on remote biometric identification is necessary in public space is necessary. Brussels (2021)
12. Gaier, A., Ha, D.: Weight agnostic neural networks. In: 32nd Conference on Neural Processing Systems (2019)
13. Hao, K.: We analyzed 16,625 papers to figure out where AI is headed next. MIT Technology Review (2019)
14. Husserl, E.: Ideas: General Introduction to Pure Phenomenology. W. R. Boyce Gibson (trans.), Routledge, Oxford (2012)
15. Ihde, D.: Consequences of Phenomenology. State University of New York Press, New York (1986)
16. Kuhn, T.: The Structure of Scientific Revolutions, 3rd edn. The University of Chicago Press, Chicago (1996)
17. Nagel, T.: What is it like to be a bat? Philos. Rev. **83**, 435–450 (1974)
18. Popper, K.: The Logic of Scientific Discovery. Hutchinson, London (1959)
19. Solomonoff, R.: The time scale of artificial intelligence: reflections on social effects. Hum. Syst. Manag. **5**, 149–153 (1985)
20. Sytsma, J., Ozdemir, E.: No problem: evidence that problem intuitions are not widespread. J. Conscious. Stud. **26**, 241–256 (2019)