Institutionen för idé- och samhällsstudier

Umeå Universitet

How smart is AI?

Leonard Loos

Magisteruppsats i Filosofi

Handledare: Peter Melander

UMEÅ UNIVERSITET

Umeå Universitet, 901 87 Umeå, Sverige

Abstract

The reemergence of artificial intelligence during the last 30 years has introduced several forms of weak AI to our everyday lives, be it in our smartphones or how the weather is predicted. Modern approaches to AI, using methods like neural networks and machine learning, also feel confident about creating strong AI, intelligence that is human-like or superior to humans. In this thesis, I discover the philosophical premises of artificial intelligence, how the research program views the mind and what implications this has for the form of intelligence that is being constructed. Furthermore, I derive at several criteria that need to be met to qualify a system as intelligent. To cover this rather wide field, the works of Hubert Dreyfus, an early commentator on AI, and David Chalmers, one of the most widely read philosophers of mind, are interrogated about their views on human intelligence and how such a theory relates to the possibility of intelligent machines.

Key terms: Artificial Intelligence, Neural Networks, GOFAI, Phenomenology, Philosophy of Mind, Materialism, Formalism, Dualism, Computationalism, Chalmers, Dreyfus

Index

Introduction

The development of artificial intelligence in machines has seen plenty of ups and downs in the enthusiasm about what AI can do and when it will be able to do it. After the early initial success of what is today known as GOFAI (good old fashion artificial intelligence), the hype soon faded when the project failed to produce the expected complex behavior that would eventually qualify AI as moving in on human intelligence[1]. Rather than creating strong AI, the project which started in the 1950s got stuck with low forms of weak AI. The past thirty years then have seen a resurgence of artificial intelligence, both due to new promising programming approaches, including buzzwords such as neural networks and machine learning, but also due to the development of hardware that is able to process the required workload. Today, forms of weak AI, the calculative processing of large amounts of data, have become part of our everyday lives; we carry AI in our smartphones, watches and digital wristbands while algorithms calculate which products suit our preferences and what news may interest us. This recent technical development has respawned a massive philosophic and scientific inquiry into the field of AI. The research project not only promises to serve human needs and to predict our future using extensive sets of data; it is again optimistic about creating strong AI, that is, intelligence that is human, human-like or superior to human intelligence, superintelligent[2].

In the light of these technical developments it becomes crucial to understand what exactly is entailed by intelligence in machines. This in turn begs the more basic questions of what it is to be intelligent and what enables intelligence. The philosophical question of how to define intelligence not only helps us to determine how machines need to develop to resemble human intelligence, but it may also serve to pinpoint the kind of intelligence that we are either already developing or the future machine that we want to construct[3]. Contemplating this extremely developed future, commentators like Nick Bostrom present us with several scenarios in which strong AI may either do much good for us, but if not guided correctly may also act as a malicious agent with disastrous consequences for human civilization. An understanding as suggested above is thus critical to carefully defining how machines ought to shape our future and how we make sure that the systems do what we want them to do.

[1] Frankish & Ramsey, *Franklin*, p. 15, 30; Russell & Norvig, p. 16, 28, 1020, 1026
[2] Frankish & Ramsey, *Franklin*, p. 28; Bostrom, p. 63
[3] Dreyfus, p. 79

The task ahead

The aim of this paper is to get a grip on the feasibility of strong AI. By looking at both the philosophical critique of the materialist and functionalist premises of AI, but also the possible roads which could lead to artificially creating human(-like) intelligence, I will argue that current AI approaches cannot legitimately be said to produce intelligence if compared to human intelligence, and that to move in on human intelligence machines would need to physically become like humans. If human and machine intelligence turn out to be radically different and the AI project is actually creating beings that share little with human beings, several questions about our future relation to machines arise.

As Keith Frankish and William Ramsey note, the realm created when AI and philosophy cross paths is utterly fruitful. On one hand, AI applies philosophy to get an understanding of the subject they are constructing, at the same time philosophers have a unique understanding of the AI subject itself by their access to theories of mind[4]. The field that I am approaching is hence already marked by the interlocking of science and philosophy. If I am to understand the current state of AI and how it may have to develop to turn into strong AI, it becomes necessary to look at this interjection of science and theories of the mind. For this purpose, I will draw on and compare the works of philosophers Hubert Dreyfus and David Chalmers.

The work of Dreyfus marks one of the earliest criticisms of the AI project, culminating in his early book *What computers can't do* in 1979. The later editions, renamed as *What computers still can't do*, keep the same basic criticism while considering the evolving AI research through the years. The last edition from 1992 hence also considers the neural networks research, which is still the central theme in AI thirty years later[5]. Dreyfus' publication together with Charles Taylor, *Retrieving Realism* in 2015, reflects his basic definition of the human mind and will help us to update Dreyfus' earlier work into our present time.

Another clarification needs to be done on Dreyfus' work. Especially his early work is strongly influenced by Martin Heidegger, who famously denied that the world is disclosed to humans via inner representations. Another strong influence is found in Maurice Merleau-Ponty who on one hand derived his thought from Heidegger, but there is also a lively debate on to what extent they differ. Merleau-Ponty puts explicit focus on the role that the human body plays for coping with the world: it is to him our embodied openness towards the world which is the starting point for experience[6]. This embodied subjectivism is arguably different from

[4] Frankish & Ramsey, p. 3; Russell & Norvig, p. 1021

[5] Dreyfus, p. 4

[6] Boehm, *Merleau-Ponty*, p. 45; Aho, p. 19

Heidegger who instead saw *Dasein* as the prime source for intelligibility, that is, the existential background of human life which we all share[7]. However, as Kevin Aho notes, Heidegger did not dismiss or ignore the body as an essential part of human existence and intelligence, but he saw it as secondary to and intelligible only on the basis of Dasein[8]. This difference between the two pillars of Dreyfus' work is to be noted at this point, but it will not be investigated further, for it arguably has little value to my object of interest in this paper and since both positions do not necessarily contradict each other, we can safely proceed. Furthermore Dreyfus, who has spent a lifetime dealing with both philosophers, combines the thought of both philosophers in his work on AI, which further strengthens our belief that the debate on Heidegger vs. Merleau-Ponty is not necessary to develop in my context.

Chalmers then will provide a somewhat opposing stance to Dreyfus. He is not only relevant by being one of the most widely read current philosophers on the philosophy of mind but has also been keen on applying his thought to the AI context. Furthermore, like Dreyfus he denies the fundamental materialist view on the mind and seeks an alternative approach. His renowned work *The Conscious Mind* from 1997 will be used to present the main themes of his theory, while the follow-up publication *The Character of Consciousness* from 2010 fills in some of the gaps and provides clarifications.

My analysis of the feasibility of strong AI is hence limited to the views of the two authors presented above, who qualify by their work being aimed at the interjection of philosophy of mind and AI science. Interestingly, both share some common ground on the fundamental premises of the scientific AI project, while strongly disagreeing about how to pursue the goal of strong AI. Of course, their positions are not exhaustive of the field and there will be many other views on AI and mind which one could investigate alternatively. Furthermore, I am combining the continental (or theoretical or phenomenological) philosophy of Dreyfus with the analytic approach of Chalmers, a procedure which carries with it some blurriness with regards to approaches, problems and definitions. Another challenge is posed by the scope of both philosophers. Chalmers is mainly concerned with consciousness and its core element of subjective experience while Dreyfus views consciousness as merely secondary to our basic coping with the world. Again, since both have different scopes and come from different traditions, a challenge will be to form and keep the coherence between both. Our Archimedean point will here be the way in which both authors present their findings as premises for intelligence.

[7] Low, p. 3
[8] Aho, p. 20

Related to machine intelligence is also the theme of emotions in machines[9]. It is common to set emotions in direct connection with intelligence and indeed, the philosophical tradition to which Dreyfus adheres has explicitly defined emotions and moods as defining of intelligence, consciousness and fundamental human psychology. However, discussing this topic would blow my inquiry out of proportion and will hence not be included. It also seems a too important issue to be handled briefly or as a side note. However, this should not bother the reader too much; according to AI veterans Stuart Russell and Peter Norvig, the real elephant in the room of AI is related to consciousness and experience, this is what I will focus on[10]. Another element of human intelligence is that of creativity and both Heidegger and Merleau-Ponty have had a great deal of things to say about it. Arguably, showing that human creativity is different from the processes going on in machines would lead to a similar point that I am arguing for, human intelligence being radically different than machine intelligence. However, this is a different, though interesting and valid debate. What I am instead focusing on is the practical coping with everyday situations. This focus seems valid since this is the realm which AI development is mostly concerned with, compare for instance how Google's DeepMind learns to walk or to play music, which seems to have little to do with human creativity[11]. Furthermore, there seems to be no clear point at which coping with the world and creativity can be separated. Relating back to the debate between Heidegger and Merleau-Ponty, the premises of creating something new is necessarily determined at the point where experience happens. Focusing on coping hence arguably captures a function that is prior to creativity and narrows down my scope by asking if what computers actually can do is enough to label them intelligent.

Another word must be said about the AI field itself. Though the author of this paper is himself working within Software Development, I in no way want to claim that my presentation of AI, and especially the current AI development, is complete. Rather, I rely on the philosophers' observations supported by AI literature to narrow the scope. I welcome any discussion about approaches to AI programming which may be fundamentally different than presented in this text, while remaining optimistic that the views that I am describing are inclusive enough to hold their own against current AI approaches.

[9] Frankish & Ramsey, *Scheutz*, p. 247
[10] Russell & Norvig, p. 1033
[11] https://www.youtube.com/watch?v=gn4nRCC9TwQ, https://www.youtube.com/watch?v=0ZE1bfPtvZo

Act I: Premises for machine intelligence

AI is not concerned with just any project; it tries to create something that resembles, is equal to, or superior to the human mind. Thus, the project necessarily has a particular view of how the mind works which enables programmers and researchers to develop their projects along certain premises. In what follows, the two commentators of AI, Hubert Dreyfus and David Chalmers, present their take on AI's philosophical heritage, which to both underlies the science of the human mind and behavior and has directly shaped the way the AI project approaches mind and intelligence, thereby determining the kind of artificial intelligence that is being built.

Dreyfus' genealogy of AI

The premises of AI are a lot older than the computer research program that started in the 20th century. What Dreyfus and Taylor label the *computer model*, which underlies AI research, is the culmination of a specific way of thinking about the mind and its relation to the world. This tradition has its roots in Plato, where the rationalist view of the mind was first defined[12]. Plato held that intelligence was related to the degree to which a person acts according to ideal ideas, hence his ability to grasp the truth of things[13]. This idea developed further over time, Kant would for instance later hold that platonic ideas are actually rules[14]. By the 20th century Edmund Husserl developed this tradition into concepts being hierarchies of rules.

> *Kant had a new idea as to how the mind worked. He held that all concepts were really rules. For example, the concept for dog is something like the rule: If it has four legs, barks, and wags its tail, then it's a dog* […] *Husserl, who can be regarded as the father of the information-processing model of the mind, argued that concepts were hierarchies of rules, rules which contained other rules under them. For example, the rule for recognizing dogs contained a subrule for recognizing tails.*[15]

Thus, gaining ever more complexity over time, finally the formalist approach emerged which holds that knowledge can be broken down into objective, static and all covering rules. Formalism also includes the idea of the mind containing representations of the world, symbols or rules that correspond to fundamental features of the world[16]. This view will be discussed further below, however, it is important to note that the approach amounts to the idea that if one could gather all the elementary symbols and rules that philosophers supposed were in the mind and store them in a computer, this would enable us to recreate the human mind

[12] Dreyfus, p. 177
[13] Plato, p. 283-285
[14] Kant, p. 189
[15] Dreyfus & Dreyfus, p. 4
[16] Frankish & Ramsey, *Arkoudas & Bringsjord*, p. 43

artificially[17]. This tradition attracted the attention of the AI research program, since this is exactly how computers work: a computer has a certain amount of rules stored in its database which reflects the world within which it interacts[18].

Dualist sorting

The platonic idea already implies a mental-physical distinction, a dualism. What Dreyfus and Taylor call *dualist sorting* amounts to the idea that there are bodily and extended things which are separated from mental, non-extended things. As Dreyfus notes, the infamous mind-body problem, which also applies to AI, is a direct result of the dualist idea, but it is only problematic precisely because one has already accepted a dualist distinction[19]. As noted, Descartes thought that the on the opposite side of the objective outer world, the inner mind contains representations of the outer world: what we hence need is a correct description of this world to get to truth. Following this Cartesian approach, a developer programs what the machine ought to know into its database, or in the case of machine learning, the machine is merely to some degree conditioned and then uses autonomous statistical learning. But how does the rationalist get the ideal ideas into the mind of the human subjects? Like AI, its philosophical foundation needs a mediator. Rationalists such as Descartes and Leibniz thought of the mind as defined by its capacity to form representations of external reality[20]. Dreyfus and Taylor call this idea *mediational* because it holds that in knowledge, I have contact with outer reality, but I get this only through some inner states[21]. *These representations are theories of the domains in question, representing the fixed, context-free features of a domain* and hence mediating its intelligibility[22].

Another element of cartesian dualism, which we noted in Plato, is that it idealizes the human subject as a detached observer: we need a view without emotions, needs and commitment to grasp the truths of things, in other words, we need to get rid of subjectivity to get to truth. This view is embedded in the history of thought and science, starting with the Platonic ideas to centuries later develop into the materialist turn. In this later view our knowledge of the external world comes in through receptors and by neural processes, with mental phenomenon

[17] Dreyfus, p. 3
[18] Dreyfus & Dreyfus, p. 4; Frankish & Ramsey, *Arkoudas & Bringsjord*, p. 36
[19] Dreyfus & Taylor, p. 12; Russell & Norvig, p. 1027
[20] Dreyfus & Taylor, p. 2
[21] Ibid, p. 2
[22] Dreyfus, p. xvii

of thinking and knowledge being reduced to qualia and brain-states, thereby reaffirming the dualist sorting, but now in biological terms[23].

Failures of the computational view

Dreyfus' critique on the premises of AI relies on the formalist conception of knowledge, the dualist division between mind and body, the neglect of subjectivity and the reduction of mind into physical processes. In what follows, Dreyfus will explain just what problems for AI follow from these positions and why this view on the human mind is not a feasible position to hold.

General intelligence

To Dreyfus, intelligence requires understanding which in turn sums up the human ability to skillfully interact with the material world[24]. This flexible coping with the world in all its facets is what is called *general intelligence*. This general intelligence enables humans to not merely limit their knowledge to one specific domain, we are able to apply knowledge in various and a steadily increasing number of scenarios. On the opposite side, current developments within AI tell us about the ever more heightened intelligence in machines: the computer program *Deep Blue* beats grandmasters of chess and *AlphaGo* is able to beat grandmasters at even more complex games[25]. However, none of these programs are able to apply their specific abilities in chess and other games in other domains. Their ability is not general but domain-specific[26]. Thus, the general intelligence that humans possess has been and still is one of the greatest challenges to AI. AI projects are usually built within a restricted environment, computer games are a good example where the machine only needs to interact with a predefined world to which it can be either programmed or can learn to interact with. The ontological difference is easy to see; human subjects do not interact with a limited reality but are able to cope with the real world in all its rich and changing facets. Hence, there is a great difference between domain-specific knowledge, where relevance can be decided in advance, and general intelligence, where relevance is not given[27]. This is what differs between

[23] Dreyfus & Taylor, p. 3, 11; Russell & Norvig, p. 1028
[24] Dreyfus, p. 3
[25] Frankish & Ramsey, *Franklin*, p. 22; https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html
[26] Ibid, p. 16
[27] Dreyfus, p. 33

a game of chess or a computer game, where computers have gotten extremely powerful, and human everyday living, where computers still fail miserably.

Relevance

The relevance of things to be grasped is hence one major issue for AI if relevance cannot be defined in advance[28]. In our everyday lives we as humans are involved in various sub-worlds (the world of the university, office, fitness studio, etc.). But as Dreyfus notes, these sub-worlds are not separate domains but instead are all part of a whole everyday world[29]. This establishes the problem of relevance for AI, since if reality cannot be split into clear cut frames, the software runs into trouble when it tries to interpret what is in the situation. To exemplify this problem, Dreyfus tells a story which computers have been unable to interpret correctly: two kids are on their way to a boy's house to give him a gift. On the way one of them notices that the boy already possesses the same toy that they want to give him and concludes that the boy will make them take it back. This story, which seems rather simple and even very young children are able to understand what is going on here, has been hard for machines to interpret in the way that makes sense to humans. The main problem for machines has been that the grammatical structure of the example, which does not clearly tell us what is meant by *taking it back.* Are we referring to the old or the new toy? Furthermore, it is clear to us that the two kids are on their way to a birthday party, something that is not stated explicitly in the example and which a machine has no reason to believe. Now, since the organization of computer representation into matching categories, frames, presupposes static meaning and a clear cut context, the computer's inability to correctly interpret what is going on in this story implies that human brains work in a different way than using frame which contain the matching rules and truths (in this case about gifts and how to give them)[30].

Performance problems

One could of course argue that relevance is merely a question of amount of data. With enough data, the idea of frames could still work. Approaches like the CYC project, the world's longest-lived AI system, represent just this approach of creating common sense by gathering massive amounts of data[31]. Yet if one tried to store all the facts from our rather simple everyday human example about returning a gift, it seems that there are perhaps indefinitely

[28] Frankish & Ramsey, *Arkoudas & Bringsjord*, p. 47
[29] Dreyfus, p. 14
[30] Ibid, p. xxiii; Frankish & Ramey, *Arkoudas & Bringsjord*, p. 50
[31] https://www.forbes.com/sites/cognitiveworld/2019/02/04/aint-nuthin-so-non-common-as-common-sense

many reasons for taking back a present. Furthermore, each of these facts requires further facts to be understood, for even if we focus on merely one fact, it will hardly be a proposition that is always true. For instance, a rule like *bring back gift because one does not want two of the same* would not apply for marbles or euro bills[32]. So a database would have to contain an account of all possible exceptions. And even if one listed all exceptions, there are situations which allow an exception to this exception. So, we also need all the context of all possible situations. Not only philosophically this is problematic, but if all the possible meanings, which may be infinite, would need to be stored together with all contexts possible, an infinite mass of data needs to be stored, making the machine's queries more complicated and arguably unreliable[33]. Not only would an enormous amount of hardware be necessary for storing this data, but the processes of searching through indexes and calculating the fitting response would crave an utterly powerful machine to do this as quick as a human does. This suggests that humans use forms of storage and retrieval that are utterly different from how machines do it[34]. The current approach hence seems to promise a powerful AI in an extensive array of fields, but this is far from the ability of a general intelligence.

Similarities

Regardless of the problem of infinite meaning and context, how does a computer find good mappings of what bits of data belong together? One common approach in AI is to map things according to their similarities. This is for example done in the image recognition which most of us experience every day when doing a google image search, and if you use Facebook, AI scans your images for repulsive content or perhaps to make predictions on which ads to show you. As Dreyfus however notes, the world has no intrinsic similarities, humans only experience similarities because meaning relies on human context[35]. *Fundamentally everything is similar to everything, so why should a program believe that two specific things ought to be compared*[36]? A Software can compare the pixel of picture A to picture B, to see to which extent they are similar, using objective tools as where the largest and most frequent similarities and differences are. But this is not the way humans do it and AI has in many cases failed to make the type of comparisons that a human would immediately recognize[37].

---

[32] Dreyfus, p. 59
[33] Ibid, p. xxi
[34] Ibid, p. 72
[35] https://youtu.be/LG5nCkcZm-g?t=5869
[36] Dreyfus, p. xxvi
[37] Ibid, p. xxvi, xxviii

Action without rules

Again, for now ignoring the problem of relevance and similarities, another related objection of Dreyfus regards how humans and machines know how to interact with the world. A common assumption is that we call upon rules when engaging with the world. We start by following rules which teach us to tie our shoelaces for instance. Then with time the rules become unconscious, but we still follow them unconsciously[38]. That is a common definition of learning, a model that AI research also accepts. Against this assumption, Dreyfus holds that when you really look at how people get from learning something to doing it automatically, it is not the case that the chess grandmaster calculates through all the possible moves he has learned and could do. Thus, there is no reason to think […] *that the rules that play a role in the acquisition of a skill play a role in its later application*[39]. Instead, a chess player initially needs rules for learning the game, but with experience the rules disappear, and the chess grandmaster just simply knows what to do. Yet, humans do call upon once memorized rules in some instances, namely when our interaction with the world fails. When the shoelaces that are being tied are too thick for instance, we need to return to the rules we once learned and reconsider them. Important is hence that the act of learning itself has more to do with us being involved in everyday interactions and not with storing, reproducing and calculating facts and beliefs as a machine would do it[40].

Summary

Dreyfus' critique of the AI premises results in the suggestion that our coping with the world is radically different from the way machines handle a situation. While machines make use of programmed or learned rules and sets of data to make more or less complex calculations, humans skip this structure and just seem to know what needs to be done. Though perhaps not as powerful in domains where knowledge can be framed, like games, humans cope with an infinite set of situations every day and handle this with ease, it hence seems unlikely that humans do this by calculating through all possibilities to find the best match. Furthermore, using the calculating strategy has not enabled AI programs to solve even very simple everyday human situations which to Dreyfus are directly known to us precisely because we are human.

[38] Ibid, p. xiii
[39] Ibid
[40] Ibid, p. 47

Chalmers' critique of materialism

After having presented Dreyfus' view on AI's ability to grasp and reproduce the human mind, I now call on David Chalmers' analytical approach to the mind to present his account of the scientific approach and what implications follow. David Chalmers is perhaps best known for differentiating between what he calls the *easy*- and the *hard problem of consciousness*. The easy problems try to specify physical mechanisms that result in mental activity, they are about the correlation of functions, descriptions of biological systems[41]. For example, an easy problem would be how to explain perception by looking at the function of the retina and brain activity[42]. Describing mental activities in this purely materialist way however does not explain how physical processes in the brain bring about subjective experience[43]. Thus, the hard problem concerns itself with the question of how physical processes relate to the experience of that there is something that it is like to be a person, to drink a cool beer or to see a red color[44]. In his widely read *The conscious mind* Chalmers sets out to present a theory of consciousness which describes exactly this *subjective quality of experience*[45].

Physical concepts of mind

Chalmers takes the existence of experience as a given; there is no question that I am subjectively conscious at this very moment. The core problem of physical concepts of mind is that they are unable to explain consciousness by their scientific methods, for example by trying to explain it by the workings of our neural system[46]. Modern cognitive science views the mind as the causal or explanatory basis for behavior; a state is mental in this sense if it plays the right sort of causal role in the production of behavior[47]. As in Dreyfus, this thought can be traced through history. Around the time when Descartes produced his dualist approach, the behaviorist movement created an objective psychological explanation with no room for consciousness, establishing the idea that psychological explanation can proceed while ignoring the phenomenal[48]. This later became part of the functionalist idea[49]. As noted, when in the 20th century computational cognitive science emerged, internal states were not regarded as having phenomenal character and were only viewed as relevant if they could explain the

[41] Chalmers (c), p. 4
[42] Ibid, p. 4, 6
[43] Ibid, p. 5
[44] Ibid, p. 104; Chalmers (b), p. 197, *Nagel*, p. 219; Russell & Norvig, p. 1028-2028
[45] Chalmers (a), p. 4
[46] Ibid, p. 4
[47] Ibid, p. 11; Chalmers (c), p. 103
[48] Chalmers (a), p. 13; Chalmers (b), p. 3; Frankish & Ramsey, *Arkoudas & Bringsjord*, p. 40
[49] Chalmers (b), *Block*, p. 94

physical causation of behavior, a mental state is in this view only characterized by its causal role[50]. This leads to mental concepts being analyzed functionally, *in terms of their actual or typical causes and effects*[51]. This structure is still valid today and when looking at current trends in AI or neurobiological sciences, approaches to consciousness have much to offer in explaining psychological phenomena and the brain processes that are correlated with consciousness. But this still does not tell us why brain processes should give rise to experience[52].

Conscious experience does not supervene on the physical

The doctrine of materialism which Chalmers refutes is based on the fundamental principle that all the positive facts about the world are globally and logically supervenient on the physical facts[53]. Supervenience describes a relation of dependency between high-level/B-properties which depend, supervene, on low-level/A-properties. In short, it entails that two instances that are similar in their A-properties are necessarily the same in their B-properties. For instance, if the fundamental physical structure of a rock is copied exactly, the appearance, density, etc., of both rocks must necessarily be the same. Vital to Chalmers' theory is that conscious experience (B) does not supervene logically on the physical (A)[54]. Logical supervenience entails that B supervenes on A logically, which to Chalmers is true of most relations between positive facts and physical elements[55]. However, there are some exceptions, to which he counts consciousness. If he can successfully argue against logical supervenience in the case of consciousness, it would be rendered a fundamental element of the world, in line with fundamental physical facts and laws of nature, which all cannot be reduced to any other element[56]. To make this case against reductive explanation, Chalmers makes several moves to show that consciousness is not logically supervenient on the physical and that all the microphysical facts in the world do not entail the facts about consciousness[57]. The most known move is his thought experiment regarding zombies.

[50] Chalmers (a), p. 14
[51] Ibid, p. 15
[52] Ibid, p. 115; Chalmers (c), p. 8
[53] Chalmers (a), p. 41
[54] Ibid, p. 71; Chalmers (c), p. 17.
[55] Chalmers (a), p. 41
[56] Ibid, p. 71, 87
[57] Ibid, p. 93

The logical possibility of zombies

In this thought experiment we are asked to consider the logical possibility of a zombie: my twin who is physically identical to me but lacking conscious experiences. Now imagine that right now I am drinking a cool beer, having a nice sensation of a cool bitter liquid pouring down my throat. My zombie twin, doing the same, is identical to me functionally: he will be

> *[…] processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be psychologically identical to me. All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions*[58].

The only difference between us is that all his functioning is fundamentally lacking conscious experience. There will be no phenomenal feel for a zombie of what it is like to drink a cool beer[59]. Hence, if the known positive facts of our functional apparatus (A) do not lead to consciousness (B), logical supervenience is not given; functional behavior can exist without consciousness and consciousness cannot be reduced to such[60].

Summary

Chalmers applies the definition of consciousness as having phenomenal properties, that there is something that it's like to see a color or drink a beer[61]. Chalmers concludes that consciousness does not logically supervene on the physical and consciousness cannot be explained reductively. If one tries to explain consciousness physically, there will always be the question of why these processes are accompanied by conscious experience. For most other phenomena in the world, which logically supervene on the physical, the physical facts entail the existence of the phenomena[62]. Physical explanation may be well suited to the explanation of structure and of function. But once we have explained all the physical structure and functions in the brain, there is one thing that remains unexplained, consciousness itself[63].

Act II: Alternative approaches

Both Dreyfus and Chalmers reject a materialist and functionalist view of the mind which to them reduces the mind to biological elements and neglects subjectivity. These premises are hence ill equipped to grasp what the mind is and how it works, with Dreyfus already laying

---

[58] Ibid, p. 94-95
[59] Ibid, p. 94-95
[60] Ibid, p. 123
[61] Chalmers (b), p. 197, *Nagel*, p. 219
[62] Chalmers (a), p. 107
[63] Ibid, p. 107; Chalmers (c), p. 108

out his related critique of AI. Moving on to Dreyfus' and Chalmers' own theory of mind, I am thereby also moving forward in defining the possible goals which the AI project needs to pursue if it is serious about creating intelligence.

Dreyfus' phenomenological approach

Dreyfus' critique of AI and its underlying philosophical premises can be summarized in the misinterpretation of a mediating and objectively calculating mind. That is, the dualism which views the outer world as being represented in the mind and the mental functions of calculating through all the stored data and rules when engaging with the world. Though this view must seem attractive to computer science, it to Dreyfus is obvious that the project must fail in producing something similar to the human mind; humans interact with the world in a radically different way and in doing so have different abilities and impediments than machines.

The role of the body in intelligent behavior

Relying on a dualist conception, the rationalist tradition leading up to AI research has supposed that the body at least to some degree can be dispensed with. The brain in the vat example for instance is to Dreyfus a problem which only occurs if one accepts a dualist view and in fact Dreyfus wants to claim that the body is indispensable for intelligence[64]: *After some attempts to program such a machine, it might become apparent that what distinguishes persons from machines is not a detached, universal, immaterial soul, but an involved, situated, material body*[65]. The idea of the body being essential to intelligence also reflects Dreyfus' position that the higher determinate logical and detached forms of intelligence are derived from and guided by global and involved lower forms[66]. Computer technology has hence been successful in simulating so-called higher rational functions, dealing well with ideal languages and abstract logical relations. But it appears that the non-formalizable form of information processing in question is possible only for beings which have direct access to the world by being directly and bodily situated in it[67]. As we stated earlier, the role of the body in intelligence is championed by Merleau-Ponty, while Heidegger did recognize the body as an important premise for meaning, but instead saw the fundamental premise of our human being in Dasein, that is, the experience of being in a world. Though much can be said about this difference, I choose to use the concept of the body as the more accessible fundamental form, but also because arguably the different approaches make little difference to my argument

[64] Dreyfus, p. 235
[65] Ibid, p. 236
[66] Ibid, p. 237
[67] Ibid

since both the body and Dasein are premises for structuring the world according to human interests. What follows in the sections below is a description of the ways in which the human body as situated in the world gives rise to intelligence.

Horizons

In our direct bodily situated access to the world, we experience the whole of it while at the same time being able to focus on its particular parts. Both dimensions are always there to us. This stands in contrast to the idea of a calculating machine and phenomenologists have pointed out that our everyday recognition of objects does not seem to operate by searching though indexes of data and rules. For example […] *in recognizing a melody, the notes get their values by being perceived as part of the melody, rather than the melody being recognized in terms of independently identified, indexed and refined notes*[68]. Humans also for instance grasp the meaning of a sentence, which is more than the sum of its letters and words, by using so called *perceptual horizons*: the outer horizon, as the father phenomenology, Edmund Husserl, called it, regards the *basic figure-ground phenomenon, necessary for there to be any perception at all*[69]. This outer horizon perceives all that is in the world and creates an indeterminate background on which those things that become prominent in experience are played out. This background lets the things that become explicit in experience show up as part of a whole, connected and coherent reality[70]. It is this outer horizon which in our chess example remains indeterminate, the chess grandmaster sees and knows the developing chess game, and yet provides the context by which the master can zero in on a specific chess move which is detached but an integrated part of the situation[71]. A computer on the other hand must take any information into account as possibly determinate, leading to enormous amounts of calculations, which the human mind just skips. A computer hence has no horizons[72].
This outer horizon describes how background information is ignored without being excluded. To zero in on some single important element of focus, Husserl uses the term inner horizon. We can view this in our everyday perception of things that we focus on, as me looking out the window at the wooden Tyrolean hut on the hill. The interplay of outer and inner horizons hence looks as follows: when we perceive an object (inner horizon) we are aware that it has more aspects than we are at the moment considering (outer horizon). Thus, in ordinary

[68] Ibid, p. 238
[69] Ibid, p. 240
[70] Ibid, p. 239
[71] Ibid, p. 240
[72] Ibid, p. 241

situations we perceive the whole object, not only what we visually can see, but even its hidden aspects that we know it possesses. We perceive a house, for example, as more than a façade; though visually merely seeing the front, we experience the house in its full dimension, with its sides and backside.

> *A machine on the other hand with no equivalent of an inner horizon would have to provide this information in the reverse order: from details to the whole. Given any aspect of an object, the machine would either pick up on its receptors or it would have to be explicitly stored in memory or counted out again when it was needed*[73].

Knowing what to do

The horizons, and meaning in itself, are made possible by our body and the skills which it develops[74]. When acquiring a new skill, like playing chess, we first act rather clumsy and uneasy while actively following sets of rules. But at a certain point we lose the rules and instead perform automatically. As already noted, at this point we do not recall any rules, which many would hold now have drifted into unconsciousness; rather, *we seem to have picked up the bodily skill which gives our behavior a new flexibility and smoothness*[75]. The same holds for acquiring the skill of perception. To take one of Merleau-Ponty's examples:

> […] *to learn to feel silk, one must learn to move or be prepared to move one's hand in a certain way and to have certain expectations. Before we acquire the appropriate skill, we experience only confused sensations. In the case of seeing, focusing, getting the right perspective, picking out certain details, all involve coordinated actions and anticipations. These bodily skills enable us not only to recognize objects in each single sense modality, but by virtue of the equivalence of our exploratory skills we can see the same object*[76].

Dreyfus hence concludes:

> […] *For a computer to do the same thing it would have to be programmed to make a specific list of the characteristics of a visual analyzed object and compare that list to an explicit list of traits recorded by moving tactical receptors over the same object. This means that there would have to be an internal model of each object in each sense modality, and that the recognition of an object seen and felt must pass through the analysis of that object in terms of common features*[77]

On the opposite side, my body enables me to by-pass this formal analysis. The bodily skills thus also include an argument against symbolic representation. A grandmaster of chess has no

[73] Ibid, p. 241-242
[74] Ibid, p. 249
[75] Ibid
[76] Ibid
[77] Ibid

rules or representations stored in his brain, rather, he simply sees what needs to be done with him experiencing the world as structured in a direct and meaningful way[78]. The phenomenologists Martin Heidegger and Maurice Merleau-Ponty, on whom Dreyfus constructs his theory, would say that *[…] objects appear to an involved participant not in isolation and with context-free properties but as things that solicit responses by their significance* (Dreyfus, xxviii). Thus, if we have vast experience, we have a direct sense of how things are done and what to expect. *This global familiarity thus enables us to respond to what is relevant and to ignore what is irrelevant without planning based on purpose-free representations of context-free facts*[79].

Things being ready-to-hand

Moving from perception to action, the same direct connectedness with the world also underlies human transcendental interaction with things. Heidegger and Merleau-Ponty both discuss the important way that our experience of a tool we are using differs from our experience of an object. A blind man who runs his hand along his cane is aware of its objective position and characteristics as weight, hardness & smoothness[80]. This is what an object is like. When he is using the cane, however, he is not aware of its position or its features. Rather, he handles the stick as was it a part of his body, this is what it is like to use a tool[81]. Heidegger originated this idea; when a carpenter is using the hammer for his work, he is not consciously aware of the hardness and dynamics of the hammer or hammering, nor is he aware of the nail he is holding. Instead, the carpenter transcends these objects[82]. *The important thing about such skill is that, though science requires that the skilled performance be described according to rules, these rules need in no way be involved in producing the performance*[83]. In fact, we are always at home and familiar with the world, are experts, and know what to do.

Being in the world

From what has been said, human intelligence relies on the unique way human beings are bodily embedded in-the-world and the special function this world serves in making orderly but non-rule-like behavior possible. Heidegger and Merleau-Ponty both believe that our conceptual thinking is embedded in everyday coping and that grasping things is possible only

[78] Ibid, p. xxviii
[79] Ibid, p. xxviii, 29
[80] Ibid, p. 252
[81] Ibid
[82] Heidegger, p. 38, 60, 69, 157
[83] Dreyfus, p. 253

for subjects taking a stance towards the world's objects as having them ready-to-hand, like canes, hammers and chess boards. The stance we take towards the world is hence fundamentally one of involvement[84]. Dreyfus and Taylor call such alternative approach *contact theory*. Where a mediational theory seeks knowledge as arising through some mediational element, contact theories give an account of knowledge as us attaining unmediated contact with reality[85]. Put bluntly, the human subject is always already fully involved in the situation, which enables one to maneuver it with bravura and without following rules[86].

Through the horizon-related structure of experience, we are in a way guided by our surroundings because we live in a preconceptual engagement with them[87]. For example, in Basketball, Michael Jordan is actively making sense of the court before him, articulating it into driving lanes, possible openings, passing lanes. This embeddedness also involves the experience of past games, perhaps even against the same player. But Jordan is grasping all of this without the benefit of concepts[88]. He is not calling on any rules or calculations and does not even seem to be aware of his choices, until he has taken the shot or passed the ball[89]. Rather, the field is already known to him and he acts brilliantly within it without the need to make all of the game's tiniest elements explicit in experience.

Human relevance

In a given situation, as in a game of chess, not all facts fall within the realm of possible relevance. Thus, in the context of a game of chess, the weight of the pieces is irrelevant. *But what counts as essential cannot be defined in advance, independently of some particular problem or some particular stage of some particular game*[90]. Now, since facts are not relevant or irrelevant in a fixed way, all facts are possibly relevant in some situation[91]. *Whatever it is that enables human beings to zero in on the relevant facts without definitely excluding others by calculation has to do with the way we are being-in-the-world*, that is, the subjective and social meaning which we get by us coping with the world[92]. Human beings are situated in such a way that what they need in order to cope with things is distributed around them where

[84] Dreyfus & Taylor, p. 35
[85] Ibid, p. 17
[86] Ibid, p. 36-37
[87] Ibid, p. 52, 72
[88] Ibid, p. 76
[89] https://youtu.be/LG5nCkcZm-g?t=1323
[90] Dreyfus, p. 257
[91] Ibid, p. 257, 259
[92] Ibid, p. 260

they need it to, not packed away and indexed in a database[93]. The human world then is prestructured in terms of human purposes and concerns in such a way that what counts as an object or is significant about an object already is a function of that concern. This cannot be matched by a computer, which can deal only with universally defined, context-free objects[94]. Our present concerns and past know-how always already determine what will be ignored, what will remain on the outer horizon of experience as possibly relevant and what will be viewed as relevant[95]. Thus, when we try to find the ultimate context-free, purpose-free elements, as Descartes and AI researchers tried, we are in effect trying to free the facts in our experience of just the organization which makes it possible to use them flexibly in coping with everyday problems[96].

This is not to deny that human beings sometimes take up isolated data and try to discover their significance by trying to fit them into a previously accumulated store of information. As noted, we do this when we are in an unfamiliar situation or when we step back to analyze something scientifically or philosophically. In Heidegger, if the mentioned hammer is too heavy for instance, this realization is not in first place a property of the hammer, but something related to one's own being in virtue of the tool being ready-to-hand[97]. Here the subject is able to focus in the situation, becoming aware in a sense, and reinterpreting what needs to be done for the flow of interaction to continue[98].
But even in these cases there must be some more general context in which we are at home. We also know what it is like to store and use data according to rules in restricted contexts. We do this when playing games for instance. But it is just because we know what it is to have to orient ourselves in a world in which we are not at home and how to model in our imagination events which have not yet taken place, that we know that we are not aware of doing this most of the time[99].

Summary

Fundamental to Dreyfus' theory is the human situation of being-in-the-world which also sets the premise for there being meaning which is both subjective and social. To be in the world also means having a body that is situated in and a part of the world. As part of the world there

[93] Ibid
[94] Dreyfus, p. 262
[95] Ibid, p. 263
[96] Ibid
[97] Heidegger, p. 154, 155, 157
[98] Ibid
[99] Dreyfus, p. 265

is no outer and inner between which mediation needs to take place, rather, we are situated in and as part of the world directly connected with it and are even able to transcend objects outside of our body, using them with excellence without having to calculate every move. Us directly knowing what to do furthermore depends on our world being prestructured by human meaning without the need to store and recall facts and rules about it.

Chalmers' naturalistic dualism

Against the materialist view, Chalmers holds that the character of our world is not exhausted by the character supplied by the physical facts; consciousness is a fundamental part of the world which is not reducible to physical mechanisms[100]. As Chalmers however notes, the phenomenal content of consciousness and psychological properties co-occur[101]. For example, the concept of perception can be taken wholly psychologically, denoting the process whereby cognitive systems are sensitive to environmental stimulation in a way such that the resulting states play a certain role in directing cognitive processes. But it can also be taken phenomenally, involving the conscious experience of what is perceived[102]. Though neither can explain the other, they do not exclude the other. These observations lead Chalmers to a form of dualism, that there are both physical and nonphysical features of the world. He observes that there appears to be a systematic dependence of conscious experience on physical structure in the cases with which we are familiar. So it remains plausible that consciousness supervenes naturally on the physical. Natural supervenience, in contrast to logical supervenience, is empirical and implies a natural co-relation between B and A but where B is not exhausted by A. It is this view, natural supervenience without logical supervenience, that Chalmers champions[103]. As he will note, this is not classical dualism with a separate realm of mental substance that exerts its own influence on physical processes. The dualism implied is instead *property dualism: conscious experience involves properties of an individual that are not entailed by the physical properties of that individual, although they may depend lawfully on those properties*[104]. The position we are left with is that consciousness arises from a physical substrate, the brain, in virtue of certain contingent laws of nature, which are not themselves implied by physical laws[105]. Since the physical and the

[100] Chalmers (c), p. xii
[101] Chalmers (a), p. 17
[102] Ibid, p. 18, 22
[103] Ibid, p. 125
[104] Chalmers (a), p. 125
[105] Ibid; Chalmers (c), p. xii

phenomenal co-occur, where there is a certain physical structure it is likely that a phenomenal structure is also in place.

How we get to content

The dualism just described is in Chalmers' later work laid out in more explicit terms and can from here on be traced throughout his concept of mind and consciousness. In *The Character of Consciousness* Chalmers introduces the term *structural coherence* to describe the intersection between outer world and inner phenomenon. On the one hand, there is the phenomenal part which is our experience: this idea of fundamental experience can here be exemplified with the *Mary the super-scientist argument*. To recall, Mary is a colorblind super scientist in color science and knows everything there is to know about colors, from a materialist point of view at least. But when Mary has a red experience for the first time, she learns something different and new, that red things cause experiences like *this*[106]. On the other side of the spectrum, Chalmers finds that awareness seems to stand for the active perception of the world. Awareness is a functional element and mediates information contents that are accessible to a central system and brought to bear in the control of behavior[107]. Thanks to awareness thus, things are directly accessible and reportable to us, while being accompanied by consciousness; they are once again coherent but not the same. In color processing for instance, every distinction between colors in our phenomenal experience has a pendant corresponding to a differentiating structure that we are aware of. Thus, experiences of color distinctions correspond to reportable information from awareness[108].

Direct access

Thus far, we have heard little that would make us think that Chalmers is not just reproducing an old dualist idea. However, he explicitly proclaims that the basic problem with dualist and materialist accounts is precisely that they make our access to consciousness mediated. Echoing Dreyfus, this sort of mediation would be appropriate if there is a gap between our core epistemic situation and the phenomena. But our access to consciousness is not mediated at all. Conscious experience lies at the center of our epistemic universe; we have access to it directly [109]. The structure of coherence, with the physical and the phenomenal co-occurring,

[106] Chalmers (a), p. 181
[107] Chalmers (c), p. 21
[108] Ibid
[109] Chalmers (a), p 196, Chalmers (c), p. 19

put in terms of the interlocking of awareness and experience, grants us direct access to the world, without a mediator.

Consciousness and Information

Information plays a large part in Chalmers' theory of consciousness. Indeed, he states that it may even be the most central part of a working theory of consciousness and it reflects his dualism in that information has both a physical and a phenomenal aspect[110]. On the phenomenal side, experience arises by virtue of its status as one aspect of information; there is something that it is like to drink a cold beer. Each such state has a phenomenal character, with phenomenal properties, qualia, that characterize what it is like to be in this state[111]. A zombie's states thus lack qualia, though the zombie can still drink beer and report that it is refreshing[112]. Qualia is the raw experience, like seeing green and tasting bitter, this is the quality of experience and is different than the mapping of what happens in a brain.

The physical aspect is found embodied in physical processing and is connected to awareness[113]. When aware of drinking a beer, the physical aspect of taste is enabled by a neuro hormonal reaction. This thus connects what we have learned earlier, that experience rises from the physical[114]. Information is embodied in the physical world but not in accordance with his overall theory, not reducible to purely material aspects.

Connecting both aspects of information is intentionality, we are intentional beings because we represent what is going on in the world. Our specific mental states, such as perceptions and thought, often have a phenomenal character and intentional content which serve to represent the world[115]. This gives rise to representationalism, the idea that physical properties correspond to our phenomenal properties[116].

Another basic aspect of information to Chalmers is that there are natural patterns of similarity and difference between phenomenal states. These patterns yield the difference structure of an information space, phenomenal states realize information states within those spaces. For example, regarding the space of simple color experiences, abstracting the patterns of similarity and difference among these experiences, we obtain an abstract information space which the phenomenal space realizes. Any given simple color experience corresponds to a

[110] Chalmers (a), p. 123, 284
[111] Chalmers (c), p. 104; Russell & Norvig, p. 1033
[112] Chalmers (c), p. 295
[113] Ibid, p. 26
[114] Ibid
[115] Chalmers (c), p. 340; Chalmers (b), p. 5
[116] Chalmers (c), p. 349, 391

specific location within this space. To find information spaces realized phenomenally, we rely on the intrinsic qualities of experiences and the structure among them, the similarity and difference relations that they bear to each other, and their intrinsic combinatorial structure[117]. Any experience will bear natural relations of similarity and difference with other experiences, so we will always be able to find information spaces into which experiences fall[118].

Summary

To Chalmers then, we are directly connected with the world with the outer elements directly being related to our subjective phenomenon of them. This is not classic dualism because there is no mediation in the traditional sense, humans have direct access to the world, which is structured by universal laws of similarity and difference, in virtue of the structure of coherence. Consciousness is the realm of experience which is different than factual knowledge. But it co-occurs with physical information. This is why Mary the super scientist learns something new when she sees a color for the first time. This experience is something related to but not reducible to physical mechanisms such as our neural structure, but they occur together.

Act III - Showdown

I now feel able to establish an account of the arguments for and against the possibility of intelligent machines, strong AI, having presented arguments against a materialist and functionalist approach to the mind and two quite different descriptions of alternative approaches. In what follows, I shall let the two constructs debate each other in the context of artificial intelligence, thereby establishing what both theories would commonly name as premises for intelligent machines and where they deviate.

Common denominators and deviating aspects

Though Dreyfus and Chalmers focus on different elements of the mind, they share some fundamental premises. Both strongly object to the materialist presumption that the mind can be reduced to physical processes. Though both agree that the mind is something that arises from our body, be it the brain, our neural structure or perceptive apparatus, a materialist description of mind is not to capture our fundamental ability to grasp and interact with the world[119]. Inevitably, such scientific approaches may tell us important aspects of the physical

[117] Chalmers (a), p. 284
[118] Ibid
[119] Ibid, p. xi; Dreyfus, p. 163

processes that accompany mind but must remain blind to mind itself. Both also deny dualism, in Chalmers' case at least in its purest form. That is, both reject the distinction between outer world and inner mind. Rather, human beings have direct access to the world and skip any mediational process which transports information from the outer world to an inner mind. It is stunning that Dreyfus' Heideggerian approach to objects, with humans transcending hammers and chess boards, resembles what Chalmers' labels the extended body/mind, that is, incorporating other objects into our subjectivity[120]. Here however the differences begin to take form. Dreyfus rejects dualist foundations all together; to him the human subject is directly geared into and a part of the world and it is a mistake to start from the premise of separating these realms; rather the world and the human subject are one and the same and should be treated as such. Dreyfus would therefore necessarily reject Chalmers' approach which favors natural- or property dualism, the co-occurrence of the physical and the phenomenological in virtue of which humans have direct access to the world. In Chalmers, information (qualia) travels between outer and inner and amounts to phenomenal content, consciousness, the difference between humans and zombies. As we have seen, Dreyfus would also reject this position since to him qualia or representations are per definition just another result of materialist and dualist thought; these concepts are only relevant if one accepts the dualist distinction of outer and inner[121]. Dreyfus would also answer Chalmers that his notion of information, related to a stimulus input, is highly ambiguous: nobody knows why the mind sees things as it does and every such mediational theory is suspiciously materialistic[122].
Chalmers holds that a basic aspect of information is that it reflects natural patterns of similarity and difference which yield the difference structure of an information space. Phenomenal states realize information states within those spaces[123]. This stands in sharp opposition to Dreyfus who holds that there are no essential differences: fundamentally everything is similar to everything and the only way we are able to coherently differentiate between objects is because objects are given meaning according to human interests[124].
On a more global level, the differences between Chalmers and Dreyfus also relate to their subject of interest. Chalmers believes that it is consciousness, subjective experience, that is the mystery of the mind which science has not been able to pinpoint or define. Though a zombie would be able to do the same things as me, it would lack mental content, phenomenal

[120] Chalmers (e), Chalmers (f)
[121] Dreyfus, p. 181
[122] Ibid, p. 180
[123] Chalmers (a), p. 284
[124] Dreyfus, p. 266

meaning. To Dreyfus on the contrary, consciousness is something secondary. It arises when we need to actively reflect upon the world. But most of the time we are not reflecting in Chalmers' sense and are instead involved in actively coping with the world without such experience playing any part. Now, these differing focus points do not exclude each other a priori. It seems feasible that even in coping with the world there are green color- and cool beer experiences which one explicitly becomes aware of. The fundamental difference is that to Chalmers everyday coping is distinct from consciousness in that it may exist in a purely functional form. In theory, a zombie could handle everyday coping without being conscious. In contrast, to Dreyfus it is instead everyday coping that is specifically human; a zombie could never do the same things as we do, what in Chalmers would be the functional acting, precisely because, relating back to what we just said about human interests, this interaction with the world presupposes the human being-in-the-world with all its facets of human relevance, motivation and transcendence. To Dreyfus hence, there can be no such zombies; either they would not act as humans or if they act as humans, this would be because they share the specific human state of being-in-the-world.

How smart is AI?

Having established the similarities and differences between Chalmers and Dreyfus, I can now apply them to the question of artificial intelligence. First however, a definition of intelligence is needed. If we want to know if and in what ways machines are able to be intelligent, we ought to know what we mean by intelligence.

What is intelligence?

We have already distinguished between domain specific intelligence and general intelligence. Domain specific intelligence is a certain expertise within a narrow field of application, like a computer- or chess game. Within such a closed domain, data can be gathered and calculated to create an enormously powerful ability. This is because there are predefined sets of rules and parameters within which the data can be used to find optimal solutions. As Dreyfus has shown, this is not how human intelligence works. For one, we do not calculate all the possible moves in a chess game like a computer does. Secondly, we are able to apply our knowledge universally, what is called general intelligence. This relates well to a philosophically common interpretation of intelligence as connected to understanding or grasping[125]. Both Plato and Aristoteles thought of intelligence as the ability to grasp fundamental facts of the world[126].

[125] Bengtsson, p. 264

[126] Plato, p. 283-285; https://plato.stanford.edu/entries/aristotle-ethics/

Eventually, intelligence would be replaced by the notion of understanding by thinkers like David Hume and John Locke[127]. As a working definition of intelligence, we can thus postulate that intelligence roughly regards the ability to grasp the world by its fundamental features, whatever they may be. Hence, AI must, in a way to be defined, be able to understand, grasping the fundamental elements of the world.

Intelligence in AI

To Dreyfus, understanding a situation is not about connecting an object to static meaning. Meaning is not natural but utterly shifting and only known to organisms which share the same type of life and priorities[128]. Intelligence is hence enabled by the body which is situated in and is part of the world. We are by having a body directly geared into the world which enables us to act intelligently without analyzing every bit of it, always doing what is required in virtue of our understanding. Sure, humans also seem to show domain specific intelligence: we recall rules and functions when our actions in the world run into an unexpected error and we need to correct our course. But since we have seen that even playing chess or go is done in a radically different way than computers do it, it is not clear that even mathematical calculation is done in the same way that calculators or AI algorithms do it. Understanding in Dreyfus is thus on one hand a purely practical notion: it regards our ability to skillfully cope with the world. On the other hand it is a reflective understanding, which, in contrast to the cartesian notion, is not purely objective but also part of the human situation of being-in-the-world.

Similarly, consciousness is usually held to also be an essential part of intelligence[129]. Chalmers' definition of intelligence and understanding can be extracted by relating to his double-punch natural dualism, the structure of coherence, that is, there are worldly things that correspond to our mental phenomenon of them. On his view, understanding is on one hand the mere functional process of data processing, of how a barley-based fluid is poured down one's throat from a glass. On the other hand, there is the phenomenon of experiencing drinking a cool bee. This is why Mary the super scientist learns something new when experiencing color for the first time. Both are due to Chalmers' double punch necessary for intelligence and understanding. Though experience carries knowledge about the world that is beyond empirical facts, these facts and the experience appear together and are both needed. As we

[127] Nidditch, Locke, p. xxii; Hume, p. x
[128] Dreyfus, p 36
[129] Frankish & Ramsey, *Robinson*, p. 67

will see, that the functional and the phenomenal belong together is also Chalmers' argument for strong AI.

Chalmers' argument for Strong AI

The zombie example is supposed to show that we can have a subject acting intelligently but without being conscious of his actions and perceptions. This is of course merely hypothetical, as a tool meant to deny reductionism. Rather, Chalmers will defend the possibility of strong AI precisely because he thinks that zombies are logically possible, but not in practice: instead, as we have seen, the functional processes of the mind and perceptive organs go hand in hand with consciousness: if a certain functional structure is in place, consciousness is likely to follow. The argument for strong AI then, again reflects Chalmers' dualism, that the physical co-occurs with the phenomenal. Relating this doctrine to intelligent machines, Chalmers places the notion of *implementation* before his dualism:

> *The strong AI thesis is cast in terms of computation, telling us that implementation of the appropriate computation suffices for consciousness*[130] […] *Informally, we say that a physical system implements a computation when the causal structure of the system mirrors the formal structure of the computation*[131].

The mentioned computations are generally specified relative to some formalist structure, like neural networks[132]. Chalmers' analysis hence is based on the idea that a system implements a computation, like a neural network, if the causal structure of the aimed at system mirrors the formal structure of the computation. Implementation is thus the nexus between abstract computations and concrete physical systems[133].

> *A physical system implements a given computation when there exists a grouping of physical states of the system into state-types and a one-to-one mapping from formal states of the computation to physical state-types, such that formal states related by an abstract state-transition relation are mapped onto physical state-types related by a corresponding causal state-transition relation*[134].

For a conscious system then, its known functional organization could be abstracted into a program x. Any physical system that implements this program x will then have a functional organization that duplicates the neuron level functional organization of the brain, for instance.

---

[130] Chalmers (c), p. 296
[131] Ibid, p. 298
[132] Ibid
[133] Chalmers (d)
[134] Chalmers (c), p. 298

Due to the co-occurrence of the physical and the phenomenal, this system is likely to have experiences indistinguishable from those associated with the brain:

> *Indeed, in an ordinary computer that implements a neuron-by-neuron simulation of my brain, there will be real causation going on between voltages in various circuits that precisely mirrors that patterns of causation between the neurons. For each neuron, there will be a memory location or location that represents the neuron, and each of these locations will be physically realized in a voltage at some physical location. It is the causal patterns among these circuits, just as it is the causal patterns among the neurons and in the brain, that is responsible for any conscious experience that arises*[135].

Organizational invariance

The system which represents such abstract causal organization, has what Chalmers names a causal topology[*136*].

> *Any system will have causal topology at a number of different levels. For the cognitive systems with which we will be concerned, the relevant level of causal topology will be a level fine enough to determine the causation of behavior. For the brain, this is probably the neural level or higher, depending on just how the brain's cognitive mechanisms function*[137].

This topology in AI rests on *organizational invariant*s. A property is an organizational invariant if *it is an invariant with respect to a causal topology: that is, of any change to the system that preserves the causal topology preserves the property*[138]. Most properties are however not organizational invariants. For example, the property of flying is not, we can move an airplane to the ground while preserving its causal topology, but it would no longer be flying[139]. *Change the features in question enough and the property in question will change*[140]. In contrast, mental properties are organizational invariants: it does not matter how we change small parts of a cognitive system, as long as we preserve its causal topology, we will preserve its mental properties[141]. This central claim can be justified by dividing mental properties into Chalmers' two basic categories: psychological or material properties, characterized by their causal role, and phenomenal properties[142]. Phenomenal properties are organizational invariants because:

---

[135] Ibid, p. 301
[136] Ibid, p. 298
[137] Ibid
[138] Ibid
[139] Ibid
[140] Ibid
[141] Ibid
[142] Ibid

> *Assume conscious experience is not organizationally invariant. Then there exist systems with the same causal topology but different conscious experiences. Let us say this is because the systems are made of different materials, such as neurons and silicon; a similar argument can be given for other sorts of differences. As the two systems have the same causal topology, we can (in principle) transform the first system into the second by making only gradual changes, such as by replacing neurons one at a time with I/O equivalent silicon chips, where the overall pattern of interaction remains the same throughout. Along the spectrum of intermediate systems, there must be two systems between which we replace less than ten percent of the system, but whose conscious experiences differ. Consider these two systems, N and S, which are identical except in that some circuit in one is neural and in the other is silicon*[143]*. If all this works, it establishes that most mental properties are organizational invariants: any two systems that share their fine-grained causal topology will share their mental properties, modulo the contribution of the environment*[144].

Hence, Chalmers' argument for Strong AI is reflective of his general theory of mind, that is, the dualist co-occurrence of the physical with the phenomenal. Furthermore, the mind is in some way a physical object and if the human mind is a consequence of physical laws, which are computable, then the human mind must be computable too. With regards to the zombie example, consciousness is distinct from but not separable from the functional processes and they both seem to occur together: hence if there is a zombie which functionally is like me, chances are that the zombie also is conscious. Hence, though we do not know much about what consciousness is and how it comes into being, the presence of the functional apparatus makes a strong case for there possibly being consciousness, and thus strong AI.

Simulation

We may be skeptical about the formalist structure amounting to consciousness. A point of critique would thus be that Chalmers' argument would arguably stay a simulation of mind, it would have no inner life and thus no true understanding[145]. As Chalmers however notes, a simulation of a phenomenon is not the same as a replication of the phenomenon. A hurricane simulated on a computer is not a real hurricane; no-one gets wet. But according to Chalmers, for some properties, simulation is replication. For example, a simulation of a system with a causal loop *is* a system with a causal loop. To Chalmers a simulation of X is an X precisely when the property of being an X is an *organizational invariant*[146]. This relates back to Chalmers' natural dualism, which enables him to view phenomenal properties as such

[143] Ibid
[144] Ibid
[145] Chalmers (a), p. 293
[146] Ibid, p. 307; Chalmers (d)

organizational invariants. It hence follows that the right sort of simulation of a system with phenomenal properties will itself have phenomenal properties, by virtue of replicating the original system's fine-grained functional organization[147]. Hence, in contrast to the computer simulation of a hurricane, a simulation of the functional system of a human being may result in consciousness.

Neural networks

AI is roughly and commonly divided into two traditions; Symbolic AI and Neural Nets[148]. Dreyfus' early critique from the 70s on regarded the inability to create a general intelligence, or Strong AI, within GOFAI, the most basic approach to AI by doing massive symbolic representation. GOFAI failed to create the type of intelligence it set out for, according to Dreyfus, exactly because it is not feasible to try to imitate human relevance by an enormous mapping of data. The current neural networks approach, with its manifestation in machine learning and the like, however, differs in some fundamental respects[149]. One need not store any symbols or rules at all in the machine. Instead, the machine has a history of training input-outputs pairs, and the network organizes itself by adjusting its many parameters so as to map inputs to outputs[150]. Neural networks have thus been very able to recognize patterns and pick out similar cases, doing this all in parallel thereby avoiding the bottleneck of serial processing[151]. But the commonsense-knowledge problem remains a problem for neural networks. There are many examples where neural networks that were given a massive amount of data to train on failed to produce the expected result due to their lack of human understanding. For instance, trying to create a software that would recognize hidden tanks in forests, neural networks were trained with extensive images of tanks and forests. What the software however learned was merely distinguishing between clouds and sun[152]. This failure was due to the images of the tanks being taken on cloudy days, while photos of the forest were taken on sunny days[153]. The software worked as designed, but it did not do what a human would have done. In the same manner, a system was built to learn to recognize the difference between dogs and wolves by being fed images of wolves and dogs. As in the tanks in forests, the network didn't learn the differences between dogs and wolves: instead, it

[147] Chalmers (a), p. 307
[148] Frankish & Ramsey, *Franklin*, p. 15
[149] Ibid, *Arkoudas & Bringsjord*, p. 52
[150] Dreyfus, p. xv
[151] Ibid, p. xxxiii
[152] Dreyfus, p. xxxvi
[153] https://medium.com/merantix/picasso-a-free-open-source-visualizer-for-cnns-d8ed3a35cfc5

recognized that the biggest difference between the images lay in the wolves usually being portrayed with snow in the background. It hence differentiated between dogs and wolves by looking at which animal as pictured together with snow.[154] This proves Dreyfus' point that human relevance is different than machine relevance with the one being a result of being-in-the-world (subjectivity and shared social meaning) and the other being purely objective. This enables Dreyfus to repeat his critique of AI, that also networks must share human understanding by sharing human being-in-the-world[155]. Only then could the network enter situations with perspective-based human-like expectations that would allow recognition of unexpected inputs[156].

Dreyfus also addresses the current concept of using reinforcement learning within machine learning, where the neural net receives a positive or negative signal regarding its action makes the system learn from experience[157]. These procedures attempt to learn about relevance by keeping track of statistics during trial and error learnings.

> *For each possibly relevant feature of a situation, the procedure keeps track of statistics on how things work when that feature takes on each of its possible values. If on the basis of these statistics the value of the feature seems to affect actions or values significantly, it is declared relevant. The situation receives an ever finer description as the set of features discovered to be relevant grows*[158].

As Dreyfus will admit, something vaguely like this is probably what the brain does. However, there are some serious problems which we already know by now: first, a feature may not be relevant to behavior on its own and the program would need to gather statistics on the relevance of combinations of features, which may be infinite, leading to the performance issues already described. Second, it is assumed that the relevance of a feature is a property of the domain. But as we have seen, a feature may be relevant in certain situations and not in others. We would therefore need to gather relevant data separately for each situation, again leading up to our known performance problem. Statistical gathering thus does not seem a practical way for computer procedures to deal the relevance-determination aspect of intelligent behavior[159].

Neural networks are to Dreyfus thus not the solution to the question of how to create strong

---

[154] https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982

[155] Dreyfus, p. xxxvi, xxxviii

[156] Ibid, p. xxxviii

[157] Ibid, p. xxxix, xli

[158] Ibid, p. xliii

[159] Ibid

AI. Yes, they do not require that one represent everything that human beings understand simply by being human. *But they encounter the other horn of the dilemma: one needs a learning device that shares enough human concerns and human structure to generalize the way humans do*[160].

Chalmers has a different take on neural networks. He does agree that computers that work on mere symbolic representation are not capable of strong AI. But the class of computational systems is much broader than this and a simulation of the brain is not a symbolic computation of the sort that Dreyfus wants to criticize[161]. Computers are utterly complex in their structure and working and with enough complexity, Chalmers believes that they can reach some level of having a mind, even in Dreyfus' sense. Chalmers further counts Dreyfus' critique to the so-called functional objections, which try to establish that computational systems could never function like cognitive systems for there are certain functional capacities that humans have that no computer could ever have. For example, it is argued that because these systems follow rules, they could not exhibit the creative or flexible behavior that humans exhibit. The argument Chalmers holds up against these objections regard the success of computational simulation of physical processes in general. It seems that we have good reason to believe that the laws of physics are computable, so that we at least ought to be able to simulate human behavior computationally[162]. And of course, if we can simulate consciousness, Chalmers believes that it in fact is consciousness. What his objection comes back to is hence the idea of structural coherence.

Summary and conclusion – of calculators, zombies and Frankenstein's Monster

The discussion leaves us with several implications on the limits and possibilities of strong AI. To Dreyfus, the kind of AI that has been developed up until now, including current robotics and learning neural networks, are little more than complex calculators. Their information processing and existential situation is fundamentally different than the way humans skillfully interact with the world. And even if such machines may become more complex with better hard- and software, their way of interacting with the world is fundamentally different from human intelligence. Relating back to our chess example, comparing the intelligence of Deep Blue and chess grandmaster Garry Kasparov is hence like comparing apples and pears: while Deep Blue uses calculations and statistical memory to play chess, Kasparov needs no such thing and thus has a mind that works completely different. This does not mean that Dreyfus

[160] Ibid, p. xlvi
[161] Chalmers (a), p. 308
[162] Ibid, p. 292

believes that strong AI is not possible: but this sort of intelligence would have to rise from something resembling the human organism, in for instance having a body and existential premises. Software alone seems ill equipped to replicate human intelligence.

Chalmers has a somewhat similar view as Dreyfus here. He holds that a specific physical system with enough complexity, whatever this complexity consists of, would through the structural coherence amount to conscious intelligence. We just need to know what this structure is. However, Chalmers thinks that intelligence does not rely on flesh and blood, but on the systems functionality, which could as well be made out of silicon. Chalmers and Dreyfus thus both agree that a complex enough machine may simulate human intelligence. Chalmers' zombie is precisely what to Dreyfus a very smart machine is like, which to him however lacks consciousness and intelligence. However, to Chalmers a complex enough formal system which resembles the functional workings of a conscious system should amount to consciousness and hence intelligence, thus no longer being a mere simulation. This distinction depends on if Chalmers' idea of property dualism and its structure of coherence holds true; if given the functional duplication of the physical human system amounts to something more than this physical structure, namely consciousness.

An outlook

From what I have said so far, several implications follow. The attempt to try to create strong AI is surely a risky thing. On one hand Dreyfus warns us that even an utterly powerful machine will not share human concern and understanding due to us having fundamentally different ways of coping with the world, found in the human embeddedness in the world, something that is not replicable via calculation. If a superintelligent machine is thus created, why should it share our interests or be able to understand what we want it to do[163]. On the other hand, Dreyfus' observations pose a challenge to current themes in AI research, but it does not offer solutions in scientific terms, but instead states an alternative story about the mind. For instance, it seems plausible that the organization of meaning is done radically different than machines do it, but how the structure behind it works we are not told. Dreyfus believes that the mind is a part of the body, but how to reproduce this structure it is up to science to find out. On the other end of the spectrum, Chalmers' arguments for strong AI which is conscious seems a lot more concrete and appealing to an analytic and scientific approach, connecting the scientific or functional with the phenomenal. This structure would to Dreyfus however need more than a neural network and the like, but also a body of flesh and

[163] Bostrom, p. 169

blood for instance. Chalmers' doctrine and argument for AI also stands and falls with the dualist structure of coherence. A suggested task for the future would be to argue for and against this position. Dreyfus presumably has a clear standpoint on this, namely that the structure of coherence is once again a constructed problem which arises first if one views the mind and the world as two distinct elements.

Apparently, this leaves us with basically three scenarios. From what I have said, the most probable scenario is that extremely capable machines will be developed in the future, they may even act like humans and in some sense become indistinguishable from humans, but they will be zombies, drinking beer without having any idea about what it is like to have a cold one when one is thirsty. Sad indeed.

Secondly, if we do by chance or plan develop (strong) artificial intelligence it is questionable if this intelligence would relate to the world as humans do. This AI may be conscious and be motivated in a different way and have a very a different being-in-the-world than humans. A machine would for instance be unlikely to worry about and be marked by the psychological and existentialist elements such as certain death and the relation between the infinite and one's finiteness. Though moods and emotions were left out of this inquiry, to thinkers like Kierkegaard (fortvivelse) or Heidegger (Furcht & Angst) these are some of the fundamental features of human being and psychology and essential to the motivation of a certain human psychology[164].

Thirdly, the most unlikely case, we may at least duplicate the human being. There is however much doubt that this can be done via software, rather sciences as neuropsychology and biology will have to do their part in this. It may not be a coincidence that Frankenstein's Monster was in fact not a machine, but a being patched together of humans.

That we will build very powerful and capable machines is no longer up for debate. But if what we have said this far holds true, there is a strong risk of creating zombie machines that act like us, but where the lights inside are out. And even if we do create intelligent machines, they must not necessarily be intelligent in a human sense. Adding to the three scenarios above, we should consider that, regardless of if the future AI is intelligent in either Dreyfus' or Chalmers' sense, it may potentially become more powerful than humans. For us to be able to define what kind of machines we want and what we want them to do it is vital that we become

[164] Kierkegaard, p. 18, 20, 23; Heidegger, p.140, 182

clear about what separates us from machines, what different types of intelligence there are and what intelligence we want to lend to machines.

Literature

Aho Kevin, – *The Missing Dialogue between Heidegger and Merleau-Ponty: On the Importance of the Zollikon Seminars*– Body and Society, 2005, https://www.academia.edu/3764336/The_Missing_Dialogue_between_Heidegger_and_Merleau-Ponty_On_the_Importance_of_the_Zollikon_Seminars (visited 31.05.2019)

Bengtsson, Jan –*Filosofilexikonet* – Forum, Borås 1991

Boehm, Gottfried – *Was ist ein Bild?* – Wilhelm Fink Verlag 2001

Bostrom, Nick - *Superintelligence: Paths, Dangers, Strategies* – Oxford University Press, Great Britain 2014

Chalmers, David J. (a) – *The Conscious Mind – In Search of a Fundamental Theory* - Oxford University Press, New York 1997

Chalmers, David J. (b) – *philosophy of mind – classical and contemporary readings* – Oxford University Press, New York 2002

Chalmers, David J. (c) – *The Character of Consciousness* – Oxford University Press, New York 2010

Chalmers, David J. (d) - *A Computational Foundation for the Study of Cognition* - School of Philosophy, Australian National University, Australia 1993/2012, http://consc.net/papers/computation.html (visited 16.04.2019)

Chalmers, David J. (e) – *Extended Cognition and Extended Consciousness* – 2017, http://consc.net/papers/excexc.pdf (visited 06.05.2019)

Chalmers, David J. (f) –*The extended mind* – 1995, http://consc.net/papers/extended.html (visited 06.05.2019)

Dreyfus, Hubert L., **Taylor**, Charles - *Retrieving Realism* – Harvard University Press, London 2015

Dreyfus, Hubert L. - *What Computers Still Can't Do: A Critique of Artificial Reason* – The MIT Press, USA 1992

Dreyfus, Huber L. & **Dreyfus**, Stuart E. – *Mind over machine – The power of human intuition and expertise in the era of the Computer* – The Free Press, New York 1988

Frankish, Keith & **Ramsey**, William M. (ed.) – *The Cambridge Handbook of Artificial Intelligence* – Cambridge University Press, Cambridge 2014

Heidegger, Martin – *Sein und Zeit*–Max Niemeyer Verlag Tübingen, Tyskland 2006

Hume, David – *A Treatise of Human Nature* – Dover, 2003

Kant, Immanuel – *Kritik der reinen Vernunft 1* – Suhrkamp Taschenbuch, Sinzheim 1974

Kierkegaard, Søren - *Sygdommen til døden* - Det lille Forlag, Helsingør 2011

Low, Douglas Beck - *Merleau-Ponty's Criticism of Heidegger* - Philosophy Today 53(Fall):273-293, September 2009, https://www.researchgate.net/publication/289991849_Merleau-Ponty's_Criticism_of_Heidegger (visited 31.05.2019)

Nidditch, P.H. ; **Locke**, John - *An Essay Concerning Human Understanding* - Oxford University Press, Oxford 1979

Platon – *Sämtliche Werke II* – Phaidon Verlag, Wien 1952

Russell, Stuart & **Norvig**, Peter – *Artificial Intelligence – A modern approach, third edition* – Pearson Education Limited, Harlow 2016

Web

https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html (visited 16.04.2019)

https://www.forbes.com/sites/cognitiveworld/2019/02/04/aint-nuthin-so-non-common-as-common-sense (visited 08.05.2019)

https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982

visited 08.05.2019)

https://medium.com/merantix/picasso-a-free-open-source-visualizer-for-cnns-d8ed3a35cfc5 (visited 08.05.2019)

https://plato.stanford.edu/entries/aristotle-ethics/ (visited 16.04.2019)

https://youtu.be/LG5nCkcZm-g (visited 16.04.2019)

https://www.youtube.com/watch?v=gn4nRCC9TwQ (visited 01.06.2019)

https://www.youtube.com/watch?v=0ZE1bfPtvZo (visited 01.06.2019)