Can Machines Become Conscious?

Candidate number: 114

ACIT4100 FALL 2021

Painting by Helen Rose. (https://bit.ly/31c2Nzz)

Contents

1. Introduction

There are some concepts in our universe that we humans sometimes spend our time contemplating. Some are weird, some existential, some remain mysteries as they have not been figured out by science yet. What is the meaning of life? Why are we here? What was there before our universe came into existence? Are there other universes out there? Are we living in a simulation? We can go on and on. I believe consciousness belongs on the list of one of the most profound, yet strange, concepts to think of and research.

Figure 1: *An Artist Resting by the Roadside* (1831-32), by Jørgen Roed. *Courtesy the National Gallery of Art, Copenhagen* [1].

When I was 15, me and my mum were in a terrible car crash. We drove in about 80 kilometers per hour when a truck slid on the ice, blocked the road and smashed into our car. I remember the seconds before the crash like it was yesterday. I knew I was going to die. I am not sure how long I was unconscious for, but when I woke up there was silence, smoke, and thankfully, my mum breathing. I kept falling in and out of consciousness in the time after the crash. Next thing I know, I'm in a helicopter. My 15 year old self was still very much in shock, and in the absurdity of it all I thought to myself: Is this how one is transported to heaven? In a goddamn helicopter? But I was on my way to the emergency, and when I realized that, the euphoria hit me. I was still alive. Still breathing. Still here. Suddenly my only goal in life was to remain conscious until I arrived at the emergency. I told myself over and over: Do not rest, do not close your eyes, you *have* to remain conscious.

I will never forget that day. In those 20 minutes in the helicopter, I got my priorities straight. If I am unconscious, it is game over. Suddenly, in the middle of all that pain and chaos, I realised that at the end of the day what matters is our conscious experience. Everything else we value and care about happens in the space of it.

I have thought a lot about consciousness ever since that day. This paper will investigate consciousness in machines and how the pursuit of making conscious machines means facing one of the deepest questions about the human experience.

Such questions are obviously big and broad and impossible to answer in one paper or even provide answers to at all. However, I will explore some

disagreements in defining and even recognizing consciousness, which is at the heart of conscious AI research. Different theories regarding consciousness will be explored, and their ethical implications studied.

This paper is structured as following: Firstly, current theories of consciousness are presented, in the quest to answer the question "What is Consciousness?". I present two prominent theories from the field of neuroscience; Integrated Information Theory (IIT) and the Global Neuronal Workspace Theory (GNW). Secondly, I present two papers from Philosophy, one by Thomas Nagel and one by David Chalmers. Thus, two different paradigms are presented: One follows the research methodologies present in Neuroscience and one which is guided by the research methodologies common to Philosophy. The result of this search, is fascinating. By looking at both theories in Neuroscience and Philosophy, we get to explore the topic of consciousness from a range of perspectives that differ in their level of abstraction. Thirdly, the theoretical frameworks will lay the foundation for a discussion into whether or not consciousness in machines is possible. Finally, the paper ends with presenting some ethical considerations revolving conscious machines.

2. What is consciousness?

The universe is filled with incredible things: Neuron stars, massive black wholes, moons orbiting moons, two trillion galaxies and the list goes on. While space is one of the most interesting things one can study, there is also this little universe within us. There are the complex structures of the mind and the seemingly impossible existence of consciousness.

Figure 2: An picture of the spiral galaxy NGC 2903, captured by the NASA/ESA Hubble Space Telescope [2].

There are different approaches to understanding our inner space. Spiritual gurus believe that by quieting the mind, and moving into the depths of our consciousness, we find eternity, peace and insights into the nature of the universe; that we are all one. Many find this mystical and unscientific. Nevertheless, many of the spiritual leaders would respond by pointing out that there is no need to *believe* in the deep experiences one can have by going into the depths of consciousness; one can experience them first hand, that is to say, subjectively.

On the other hand, there is a long list of prominent neuroscientists that argue that consciousness is something physical entirely and that at one point we will find exactly where consciousness is located within the brain and arrive at a precise mathematical definition, and if we do not, then perhaps

consciousness is just simply an illusion of the mind.

One interesting view regarding the nature of consciousness was presented by philosopher David Chalmers, as we will see later in the paper, who believes that the problem of consciousness often times gets minimized or simply ignored altogether because of how uneasily consciousness fits into our common conception of the natural world (in which most people refer to the physical world). He believes that we cannot reduce the problem of consciousness to fit into the natural world if that means we have to simplify it into something that it is not [3].

If I were to ask you right now: "Are you in there?", you intuitively *know* the answer. Furthermore, if I were to ask you: "But how do you know?", then the immediate feeling we get is we *know* because everything that we experience in the world is happening within the bounds of our consciousness. This precise and vivid, yet strange, feeling that we all share; that the lights are on, that we are in here, wide awake and present, that we indeed have consciousness, qualia and phenomenal experiences should be taken seriously by working towards a comprehensive, all-encompassing model of it. This is what Chalmers coins as *the Hard Problem of Consciousness* [3].

What this paper will explore is the notion that prior to questioning whether or not we can create conscious machines, we most probably must have a definition of consciousness that is so precise that it is programmable. Now, it is reasonable to question that because it is such a challenge defining consciousness, why would we already be thinking about conscious, self-aware or sentient machines? It is often considered that machines will eventually reach, or even exceed human intelligence. There are many dystopian sce-

narios of super intelligent, super conscious machines presented in works of literature as well as in movies and series. However, there is also a different scenario where there is an Artificial Intelligence (AI) or Artificial General Intelligence (AGI) takeover, with fatal consequences for humanity, but where the AIs do not have consciousness at all. What would this mean? Some argue that the conscious life that exist on earth (and potentially other planets too) is what brings meaning to the universe. It is how the universe experiences itself, its own glory, as opposed to just being vast, dead matter. If there are AIs without consciousness our universe would, some could argue, lose its beauty, because there would not be anyone there to witness and consciously process the information that is present. Therefore, if it is true that we will create machines that are more intelligent than we are, maybe it would make sense to also give them consciousness. To give them experience, to give them a chance to subjectively know what it feels like to be a part of our universe and our world.

Now it may be the case that consciousness is not something in need of defining or implementing into machines. Instead, it might just be something that emerges at a certain point in complex systems. In her book *God, Human, Animal, Machine*, Meghan O'Gieblyn explores the idea of emergent intelligence in AI: The notion that higher-level cognitive behavior or capacities can spontaneously appear (through self-organization and emergence) in machines without first having been programmed [4]. She looks at the work of Rodney Brooks, former director at the MIT Artificial Intelligence Lab, who focused on establishing an "embodied intelligence" approach to robotics in which the body of an intelligent system plays an important role [4]. The

goal was to recreate the conditions in which human evolution takes place. At MIT, Brooks and his team researched whether or not simple, primitive rules could give rise to complex behaviors such as conscious phenomena, which was one of Brooks predictions.[4]. For a long time, critics of artificial general has argued that we will not achieve such levels of intelligence, because we have not yet arrived at a complete theory of the brain. Nevertheless, emergence in nature suggests that complex systems can, in spontaneous ways, self-organize without having been programmed that way. Order can emerge from chaos. The dream then, is that perhaps simple rules organized in the right ways, can give rise to consciousness - as a biproduct of complexity. [4].

We have now seen that there are different ways of approaching the development of consciousness in machines. One way, is to start with a comprehensive definition of consciousness, and them program it into machines. The other way is suggesting that maybe we do not have to focus on how to program consciousness, but how to program the specific rules that allows for consciousness to emerge. The following sections of this paper will investigate both approaches, based on theories that take on different methods.

What a time to be alive. We have not yet established an all-encompassing theory of consciousness. Artificial General Intelligence is not solved. Anything is possible, and exactly because of that, it is all the more important that we keep thinking about these issues regarding consciousness.

2.1. Current Theories of Consciousness

One intriguing, as well as slightly frustrating, aspect of the study of consciousness, is the lack of consensus among researchers. A spectrum of ideas are discussed, ranging from those who believe that consciousness is an illu-

sion, to those who thinks that it pervades all things. Some hope to see it "reduced to the underlying biology of neurons firing; others say that it is an irreducibly holistic phenomenon" [5]. As such, the search to establish a theory that explains the neural basis of consciousness remains one of the greatest and most fundamental challenges in modern neuroscience. There has been developed a significant amount of highly sophisticated models and theories attempting to define and formalize how consciousness is implemented in the brain. Various interdisciplinary approaches has been made, such as exploiting insights from neuroscience, computer science, philosophy and psychology [5]. Among them are two important, major theories that are competing in the field. These two are: The Integrated Information Theory (IIT) and The Global Neuronal Workspace (GNW) theory. The difference, as we will see, generally lies in their "level of conceptual abstraction and anatomical specificity" [6].

The following sections will explore these two theories, and therefore this part of the paper will serve as a deep-dive into some complicated ideas. They are state of the art theories regarding consciousness from a neuroscientific perspective. However, please follow along, these ideas will lay the foundation for a further, more high level discussion of the topic at hand. Furthermore, these ideas will hopefully be intriguing and thought-provoking, as they attempt to uncover consciousness itself.

2.1.1. Integrated Information Theory (IIT)

The Integrated Information Theory was initially proposed by Tononi in 2004 [7]. What makes the IIT fascinating and compelling is that it does not in fact consider "particular brain areas or temporal profiles" [6], but instead

looks at conscious systems from the perspective of information processing and architecture. The fundamental premise Tononi makes is that "to be conscious is to have an experience" [7]. To have such an experience rages from experiencing the warmth of the sun on a hot summer afternoon, or the experience of an internal fear rising when something terrifying happens, or even the experience of having no experience at all, what is called a state of "blank mind" that can be achieved through meditation and certain yoga practices. Tononi wanted to extract the essential features and qualities that defines experiences. His search resulted in the following five axioms: That they are subjective, structured, specific, unified and definitive. Subjective refers to the idea that the experience exists for the conscious entity only. Structured points at how their contents have a relation to one another, such as *the orange carpet is on the ground*. By specific, it means that it can be formalized as being something concrete and not all things at once; so for example could we say that *the sun is yellow, not green*. Unified refers to how conscious entities experiences only one thing at a time and finally, definitive points to how all experiences have bounds to what it contains. [5]. As such, these features make up, according to Tononi and Koch, the properties of a system that contains some level of consciousness. The IIT therefore poses a prediction that the neural networks that allow for consciousness to occur, are highly interconnected, in the sense that they integrate various components of a state into an experience that is unified [6].

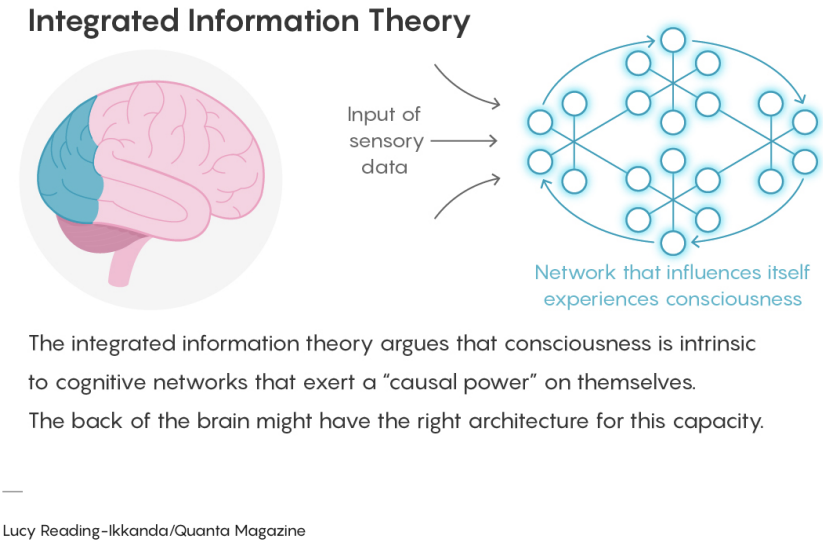

Figure 3: A visual representation of the Integrated Information Theory (IIT) model [5].

An important advantage of this theory is that it "provides a mathematical metric of irreducibility (or integration), Φ, that can be related to the level of consciousness. Proponents of IIT point to its explanatory power: for instance, it can explain why the cortex is capable of producing conscious experience while the cerebellum is not, even though the cerebellum possesses up to four times more neurons" [8][6].

2.1.2. Global Neuronal Workspace Theory (GNW)

The GNW theory, much in difference to IIT, was based on neural imaging studies in humans and primates derived empirically from EEG (which measures electrical activity on the scalp). These studies found that:

When a stimulus is presented but not consciously perceived, ac-

tivation can be seen mainly in the associated primary sensory cortices. When the stimulus is consciously perceived, however, activation in primary cortical areas is followed by a delayed "neural ignition" in which a sustained wave of activity propagates across prefrontal and parietal association cortices [6].

What this theory suggests, is that essential information flows through different regions in the brain to various subsystems for use in processes like memory, reporting and decision-making [6]. The Global Neuronal Workspace Theory is an extended version of the more fundamental theory, the Global Workspace Theory (GWT) which essentially proposes that conscious thought and behavior arises when information is stored in a "global workspace" inside the brain, and is then broadcasted to subregions of the brain associated with a particular task [5].

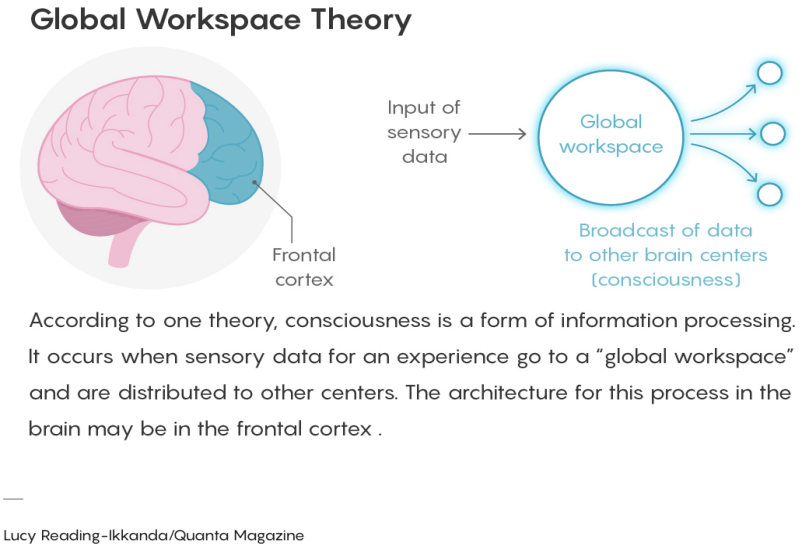

Figure 4: A visual representation of the Global Neuronal Workspace (GNW)

model [5].

The main difference between IIT and GNW, then, is that the IIT has its primary focus on "abstract connectivity and information-processing structure, GNW proposes a concrete spatiotemporal locus for conscious processes" [6]. The next part of the paper will go on to investigate some philosophical viewpoints regarding consciousness. Both viewpoints are highly influential in modern discussions of consciousness.

2.2. Philosophical Views on Consciousness

When it comes to matters of the mind and the conscious, philosophy certainly has a long and rich history. Given the enormous amount of books and papers that has been written on the topic of consciousness in philosophy, it would be an impossible task to provide here a historical overview of all of it. However, what this paper will do, is to focus on two renowned and celebrated papers that addresses the topic of consciousness:

- *What Is It Like to Be a Bat?* by Thomas Nagel
- *Consciousness and Its Place in Nature* by David J. Chalmers

The reason why the first paper is so relevant for this paper, is that it proposes that when we try to understand consciousness, we will always be somewhat limited and confined by *our* idea and conception of our own conscious states. When Thomas Nagel asks "What is it like to be a bat *for a bat?*"[9], it begs the same question that is relevant for consciousness in machines: What is it like to be a machine *for a machine?*. In the second

paper, Consciousness and Its Place in Nature, David Chalmers asks profound questions regarding the different possibilities of what consciousness can be. He wants us to contemplate the following: What if consciousness cannot in fact be reduced to matters of the brain? What if it also concerns our view of the natural world and that in order to get a complete theory of consciousness it *includes* that we also expand our view of nature? [10]

2.2.1. What Is It Like to Be a Bat?

When I was in my first year of university, enrolled in Computer Science, I read obsessively about Artificial Intelligence. Additionally, I had a strong intuition that if I were to understand artificial intelligence, I would have to not only understand how the brain works, but I would also have to indulge myself into other disciplines, such as philosophy and psychology. I enrolled in a class that was a specialization course for students in philosophy, called "Metaphysics and Philosophy of Mind". In the beginning of this course, we were introduced to this paper, *What Is It Like to Be a Bat?* by Thomas Nagel. When I read it, I was shaken to the core. Why are nobody talking about this? I am aware now that my knowledge then was limited, because philosophers all over the world are indeed talking about it, but still. This paper presented such a fundamental problem, it shed light on why it may the case, that despite of our deep quest to understand all matters in the world, we might be limited by the very thing that provides so much richness to our lives- our consciousness. As such, I am very excited to present it here, as I believe it should be mentioned in all conversations regarding conscious machines.

Figure 5: *Flying bat Painting* by Daniela Vasileva (https://bit.ly/3xzRv4k)

In the beginning of Thomas Nagel's paper *What Is It Like to Be a Bat*, he says:

> Consciousness is what makes the mind-body problem really intractable. Perhaps that is why current discussions of the problem give it little attention or get it obviously wrong. The recent wave of reductionist euphoria has produced several analyses of mental phenomena and mental concepts designed to explain the possibility of some variety of materialism, psychophysical identification, or reduction. But the problems dealt with are those common to this type of reduction and other types,and what makes the mind-body problem unique, and unlike the water-H2O problem or the Turing machine-IBM machine problem [...], is ignored [9].

The reason why, according to Nagel, the problem of consciousness is unique and different, is that the problem at hand concerns the subjective character of experience. Fundamentally, he suggests, an entity has conscious mental states "if and only if there is something that it is like to *be* that organism- something it is like *for* the organism" [9]. Thus, it must be distinguished from other analyses that attempts to reduce the mental.

As such, Nagel proposes "It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing" [9]. This quote begs so many questions. How can we measure consciousness in machines? Is it possible? Why it is that simulation is different from the simulated systems themselves?

Importantly, Nagel does not argue that a physical account of reality, including consciousness, is not possible, or that consciousness is not possible in machines. What he instead proposes, is that the phenomenological features of experience must be included into the physical description- they must themselves be provided a physical account. However, this is no simple task, and Nagel emphasizes this when he proposes: "Facts about what it is like to be an X are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them" [9]. Nagel illustrates what he means by investigating bats. He bases his illustrating example on the assumption that bats has some level of consciousness, in other words, experience. The reason Nagel chose bats is because they represent this *alien* form of life, with ways of perceiving the world so very different from our own. Nagel, as we saw, defines consciousness simply, it is the idea of experience;

"what is it like to be X?". The following quote is Nagel explaining exactly how much bats differ from the way in which we experience the world:

> Most bats (the microchiroptera, to be precise) perceive the external world primarily by sonar, or echolocation, detecting the reflections, from objects within range, of their own rapid, subtly modulated, high-frequency shrieks. Their brains are designed to correlate the outgoing impulses with the subsequent echoes, and the information thus acquired enables bats to make precise discriminations of distance, size, shape, motion, and texture comparable to those we make by vision. But bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine. [9] [10]

The idea of bats having a widely different subjective experience of the world to our own, suggests that our imagination can only get us so far. If one were to deeply imagine all the things listed above, how it must feel like taking in the world through sonar or how it is having webbing on our arms, one only imagines what it would be like for *us* to be like a bat. Nagel refuses this and proposes: "But that is not the question. I want to know what it is like for a *bat* to be a bat." [9]

Some might be sceptical to this line of reasoning. However, if were were to change seats with a bat, a machine or a Martian, and imagine if they were going to describe and understand what our subjective experience of the world is, we would be concerned immediately. Our intuition tells us that

the structure of their minds does not allow for them to form a conception of what our rich, varied, vivid experience of being alive is. If they concluded that there is *no such thing* as an experience of being human, we would be furious. Even if they cannot comprehend it, does not imply that our experience is not real, rich in complexity and with high levels of detail [9].

This moves us into the heart of the matter, a topic that is both deep and relevant when considering conscious machines: "the relation between facts on the one hand and conceptual schemes or systems of representation on the other" [9]. We have a given brain structure, and while we humans love to assume that if only we had an infinite amount of time, we would at last comprehend all information in the world, Nagel's questions this assumption. What if the structure of our brain and body has limitations to understanding consciousness? Not necessarily limitations in understanding our own, but understanding all forms of consciousness that takes place on earth and on other alien planets. Or as Nagel puts it: "Reflection on what it is like to be a bat seems to lead us, therefore, to the conclusion that there are facts that do not consist in the truth of propositions expressible in human language. We can be compelled to recognize the existence of such facts without being able to state or comprehend them" [9]. The reason why this line of reasoning is so connected to the mind-body problem, is that if facts about the subjective character of experience can only be accessible from one point of view- from the one having the experience, then how can this be disclosed in the physical operation of that organism? There are different kinds of facts. For example objective facts *par excellence* are facts that can be comprehended from many different viewpoints [9] [10]. Take for example the sun. The sun has an

objective character in that it can be observed and understood by even aliens without vision. In contrast, is rather challenging to understand the *objective* character of experience. Nagel asks a deeply provoking question: "After all, what would be left of what it was like to be a bat if one removed the viewpoint of the bat?" [9]. Thus we come face to face with the challenge of psychophysical reduction [9]. Reduction usually involves the pursuit of gaining more objectivity, a more precise definition of the nature of different phenomena. But experience simply does not fit the pattern. In other words, if we try to move further away from a specific viewpoint, in hopes of getting closer to greater objectivity, we are simply failing. We are not moving closer to understanding the real nature of the subjective phenomenon of experience by removing the specific viewpoint itself; it rather takes us farther away from it [9]. Furthermore, Nagel goes on to propose, somewhat radically:

> If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done. The problem is unique. If mental processes are indeed physical processes, then there is something it is like, intrinsically, to undergo certain physical processes. What it is for such a thing to be the case remains a mystery [9] [10].

What does all of this mean? Well certainly it makes a point of how even if we manage to create conscious machines, we cannot simply assume that their conscious experience will be similar, or even remotely similar, to our own. How many philosophical, scientific and ethical questions does this give rise to? An almost infinite amount, one could argue. How can we

know machines or AI programs are not already conscious, but because their phenomenal states differ so significantly from our own we are unable to detect it? What if we cannot in fact implement consciousness in machines because their structure (hardware) differ so greatly from our own? Maybe we need to focus on creating machines/robots that undergo the same evolutionary processes that humans ungergo?

2.2.2. Consciousness and Its Place in Nature

In his paper *Consciousness and Its Place in Nature*, Chalmers proposes several compelling ideas. If we come from a place were we want a concise, small mathematical definition of consciousness, Chalmers is bringing bad news. But since this paper aims to investigate the very complexity concerning consciousness, and why consciousness in machines is therefore no simple task, we have to explore it.

From the very first sentences in his paper, he is proposing sweeping ideas:

> Consciousness fits uneasily into our conception of the natural world. [...] So it seems that to find a place for consciousness within the natural order, we must either revise our conception of consciousness, or revise our conception of nature [11].

Chalmers divides the metaphysics of consciousness into six different classes, in which he labels type A to type F. A through C regards views that are of reductive nature. D through F focuses primarily on nonreductive views that requires changing or updating our concepts concerning physical ontology [11]. Before we go into these different classes, some terms needs defining.

Since Chalmers states that consciousness fits uneasily into our natural

world and presents bold claims regarding consciousness, we need a proper definition of the term before moving on. But defining consciousness is a challenge in and of itself. There are several reasons for this, one being that consciousness remains a matter we all have a direct relationship with, thus making it difficult to extract objective conclusions in similar ways to other phenomena in the world. Secondly, consciousness is one of the terms that has different definitions in the scientific literature, the philosophical literature, as well as peoples personal opinions- ranging from it being an illusion to consciousness being God. Chalmers approaches this challenge by dividing the problem of consciousness into two categories: The easy problem of consciousness and the hard problem of consciousness.

The easy problem of consciousness concerns phenomena that can be understood and reduced to scientific definitions in terms of computational or neural mechanisms. Examples include reporting information, reacting to stimuli, deliberate controlling behaviour or monitor internal states. As such, the easy problem of consciousness is solved through explaining and defining cognitive and behavioural functions. [11]

Then we have the hard problem of consciousness: the problem of experience. Here Chalmers refers to the exact same definition that Nagel uses- the idea that having a conscious experience essentially means that there *is something it is like* to be that entity. It points to the idea that we have a stream of conscious thought. There is something it is like to watch the sunset, feel water against the skin, breathe. Because Chalmers believes that the hard problem is categorically dissimilar from the easy one, that may indicate that categorically dissimilar solutions must be provided as well. The solution to

the hard problem of consciousness remains unknown, but Chalmers claims that it would "involve an account of the relation between physical processes and consciousness, explaining on the basis of natural principles how and why it is that physical processes are associated with states of experience" [11].

In the following section, the different classes of the metaphysics of consciousness will be presented.

Type-A materialism holds that there is no epistemic gap (a gap between physical and phenomenal truths) and even if there is a gap it is not too challenging to close it. In essence, the type-A materialist suggest that the hard problem of consciousness in actuality is one of the easy problems. Thus the hard problem is an illusion and all matters of the mind (including consciousness) can be explained, defined and understood in terms of cognitive and behavioural functions. Nevertheless, there is disagreements within type-A materialists. One version is eliminativism; consciousness does not exist. A different version admits that consciousness exists, but only in functional or behavioural terms (also known as analytic functionalism or logical behaviourism) [11].

According to type-B materialists, the epistemic gap exists but no ontological gap. Type-B materialists sees a kind of empirical identification between phenomenal states and physical states. Examples include the identification between genes and DNA, or H2O and water. The fundamental belief is that while H2O is not the exact same thing as water, they still point to the same concept in the physical world [11].

In the type-C materialist view, the epistemic gap is acknowledged as being very much real and present, but in principle it is manageable and possible

to close it. What this view suggests is that even if we lack understanding of the solution to the hard problem of consciousness in physical terms, there are still ways of solving the problem/close the gap all the while. The type-C materialist view is popular and held by many, including Churchland, Gulick, McGinn and Nagel. The view is appealing as the explanatory gap is fully acknowledged but suggests that there are ways to explain and understand the gap in terms of human limitations [11].

Type-D dualism, also known as interactionism, holds that "microphysics is not causally closed, and phenomenal properties play a causal role in affecting the physical world" [11]. This type of dualism consider physical states to cause phenomenal states and the other way around. This is quite similar to Descartes' substance dualism, that holds that there are separate mental and physical entities that are interacting. Finally, we have property dualism, suggesting there is only one substance with both physical and phenomenal properties that affect the world [11].

Type-E dualism, or epiphenomenalism, proposes there to be an ontological distinction when it comes to phenomenal truths and physical truths, but that phenomenal properties has no effect on the physical, put in other words, the physical realm is causally closed [11].

Finally we have the type-F monism in which "phenomenal or protophenomenal properties are located at the fundamental level of physical reality, and in a certain sense, underlie physical reality itself" [11]. Chalmers suggests that type-F monism can be thought of as both materialistic and dualistic. One could view it as a special kind of materalism if one assumes that by one refers to the underlying protophenomenal properties when referring to

physical properties. On the contrary, it is dualistic in that it views the phenomenal as ontologically fundamental and it "retains an underlying duality between structural-dispositional properties [...] and intrinsic protophenomenal properties" [11]. Chalmers himself holds this position, or at least he is most closely identifying with the type-F monism compared to the other classes.

We have now learned that there are a range of possible views on consciousness. What will follow, is a discussion of how these theories can be used when contemplating the very implementation of consciousness in machines.

3. Consciousness in machines

By looking at both the current theories of consciousness as well as philosophical contributions to discussions concerning consciousness, we have laid the groundwork for a further discussion- one that involves machines. It is time to think about the future. We are living in an exciting time, where Sci-Fi movies tries to make up scenarios and possible future worlds where AI systems/robots become conscious. There are lots of examples, including Westworld, ExMachina and Her.

Figure 6: Westworld (HBO season 3)

This idea of machines with levels of intelligence and consciousness that exceed that of humans, is intriguing, and to some, terrifying. In some Sci-Fi movies and Sci-Fi books, the robots are described to have artificial intelligence, or superintelligence, but often times it is implicit that these robots also have (at least most probably) some level of consciousness to reach that level of intelligence, as they often high-level cognition and information processing power in association with conscious perception [12].

In the field of Artificial Intelligence, there has been many victories in recent years, such as IBM Watson winning jeapordy (2011), DeepMind beating Atari (2013), AlphaGo beating world-champion Go player Lee Sedol (2016),

AlphaZero winning over masters in Chess, Go and Shogi (2017), and 2019 AlphaStar reaching the top 0.2 percent of human players in Starcraft [13].

However, are we moving closer to Artificial General Intelligence (AGI), Strong AI or Conscious AI? Will we soon have robots that dance better than us, or that make better art than we do? And finally, the central question of this paper: *Can* machines become conscious? Is it possible?

3.1. Can Machines Become Conscious?

Often in debates regarding Artificial Intelligence, *Strong AI* is considered to be systems with a more general intelligence usually thought of to include consciousness and sentience, while *Weak AI* refers to systems that performs narrow, specific tasks that are not conscious or sentient [14].

Earlier, in section 2.2, we saw the great challenges that exist when it comes to defining and explaining consciousness (formerly defined as "the hard problem of consciousness"). In the discussion of Thomas Nagel's paper *What Is It Like to Be a Bat?* we contemplated how consciousness is something that is experienced subjectively, from a first person perspective, and the challenges that follows. Now, the only way we can access artificial consciousness will be from the third-person perspective, highlighting a crucial challenge of how third-person witnesses can determine, evaluate and measure consciousness. There are different approaches to this very problem. One is to steer clear of establishing a narrow definition of machine consciousness or not define it in the first place. Some believe this to be a pragmatic approach: We agree on a broader definition of consciousness, and then move on with the research [14]. There are obvious problems with this approach, especially when it comes to measuring consciousness and the ethical implications of having a vague

definition. Other researchers focus instead on self-awareness.

In their paper, *Toward Self-Aware Robots*, Chatila et al. suggests emphasizes this:

> We still lack a genuine theory of the underlying principles and methods that would enable robots to understand their environment, to be cognizant of what they do, to take appropriate and timely initiatives, to learn from their own experience and to show that they know that they have learned and how. [...] The understanding of its environment by an agent (the agent itself and its effects on the environment included) requires its self-awareness, which actually is itself emerging as a result of this understanding and the distinction that the agent is capable to make between its own mind-body and its environment. [15]

Another approach to awareness, is focusing on adaptation from a system-level perspective and defining consciousness thereafter; that consciousness can be "regarded as a function for effective adaptation at the system-level, based on matching and organizing the individual results of the underlying parallel-processing units. This consciousness is assumed to correspond to how our mind is "aware" when making our moment to moment decisions in our daily life." [16]

Over and over, as we have seen throughout this paper, we come back to issues revolving *defining* and *categorizing* consciousness. One way of categorizing concepts of consciousness, that is perhaps especially appropriate in the context of machine consciousness, is the following: 1) A conscious (or sentient/wakeful) entity, 2) Being conscious (or aware) of something, such

as a sound 3) Conscious mental states, for example the awareness of feeling water against the skin [14].

The reason why this is mentioned, is that a fundamental disagreement that seems to be present in these discussions, has to do with whether we can compartmentalize consciousness into being several separate concepts, or if consciousness truly is one large phenomena that cannot be neither reduced nor divided into smaller parts.

The famous philosopher Ned Block separates *phenomenal consciousness*, which refers to experience, what it is like to be the entity, and *access consciousness* which points to "a mental state's availability for use by the organism, for example in reasoning and guiding behavior, and describes how a mental state is related with other mental states" [14]. Because phenomenal consciousness ("the hard problem of consciousness") is indeed so hard, many researchers believe it is better to start with solving access consciousness. As we will see in the next section, this might be where IIT and GNW comes in.

3.2. Responding to current theories of consciousness

In section 2.1.1 we looked at Integrated Information Theory (IIT) followed by section 2.1.2 where we looked at Global Neuronal Workspace theory (GNW). This section will look at both of these theories in the context of machine consciousness and attempt to answer the question: Can Machines Become Conscious?

The IIT can be seen as a promising theory of consciousness in the context of machine consciousness. While the IIT proposes an abstract way of defining and describing consciousness, and thus has not yet obtained "unambiguous validation [...], it provides one of the most detailed accounts for the emer-

gence of conscious experience from an information-processing network" [6]. In accordance to Tononi and Koch, our current approach to Artificial Neural Networks (ANNs), in which the information is flowing forward, feeding the network with input that is then passed through a mathematical function to convert the inputs to outputs, only produces "zombies". All the while such ANNs might be acting in ways that *imply* consciousness, without processing it. Many AI researchers around the world believe that AI systems eventually will become conscious, but this will not be the case unless they have hardware that allows for consciousness, according to Koch. [5]. Koch states that while computers may be able to simulate consciousness, the computer is in fact not conscious if the simulation has no causal power. A parallel can be drawn to computer games: Even if one uses physics equations to simulate gravity, gravity is not actually produced in the system [5]. This is in stark contrast to the IIT model, which instead argues that consciousness is inherent to cognitive networks that employ a "causal power" on themselves [5]. However, the IIT model requires extensive research before claims of having achieved machine consciousness can be taken seriously.

The GNW and the GWT suggests that consciousness is a form of information processing taking place when the input (sensory) data goes to a "global workspace" in which it is then broadcasted to other regions of the brain producing an effector output (see figure 6) [5]. The question is, can this theory be applied to computational models for machine consciousness? The Learning IDA (LIDA) architecture is one such computational implementation of the Global Workspace Theory with a rich architecture. While the computational instantiation of conceptual LIDA is ongoing, it is a model that

inspires future models of this kind for the research of machine consciousness and stands as a promising alternative to other models in AI. [17].

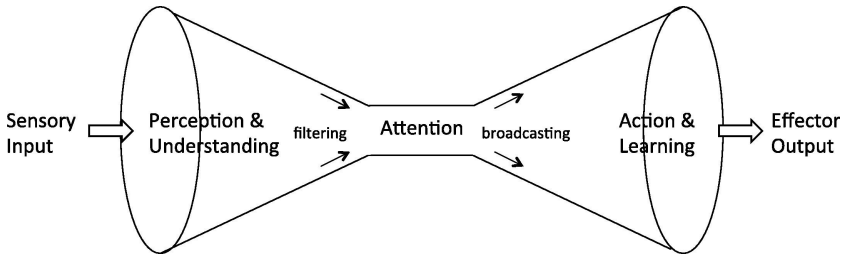

Figure 7: GWT as a bottleneck and broadcaster. Adapted from Franklin et al., 2016, fig. 1. [17]

Nevertheless, even though both the GNW and the IIT have received experimental support, the two theories are rarely compared within the same dataset. One exception is Noel et al. (2019), where IIT and GNW were their predictions were tested within the same dataset (at single-unit level). However, more research like that of Noel et al. must be conducted so that the two major theories of consciousness can be compared formally [6].

It would, however, be impossible to end this discussion here without mentioning the explanatory gap. The IIT and the GNW are unfortunately haunted by this gap; is consciousness in the brain and a primitive part of our reality, and thus can be explained in terms of other physical phenomena? *Or* does the hard problem of consciousness indeed exist; meaning matter have experiences of being matter that arises from nonconscious processes. If the latter is the case, both the IIT and the GNW fall short in explaining

that gap. How is it, that a finite number of neurons together give rise to the incredibly rich experience of drinking coffee or sensing the world around us? If we do not acknowledge the hard problem of consciousness, will we ever achieve conscious machines? And if we do acknowledge the hard problem of consciousness, will the task of creating conscious machines seem impossible to overcome?

3.3. Responding to Nagel and Chalmers

Through their respective papers, Nagel and Chalmers bring great nuance and depth to the discussions of consciousness. Nagel proposed that we have to account for the phenomenological features of experience in theories of consciousness. For consciousness in machines, that would mean defining consciousness in such a manner that a systems internal/subjective experience is a part of the system by accounting for the systems phenomenological properties of a given experience. Previously, we saw a quote by Nagel that said that facts regarding what it is to be something are very peculiar. And exactly because they are so peculiar, we may want to discard them, doubt them, ignore them or overlook the significance of them. In the previous section (3.1.1 Responding to current theories of consciousness), advanced theories of consciousness were discussed. On the one hand one could critique the IIT and the GNW for doing exactly what Nagel points out- as they "ignore" the hard problem of consciousness. However, since consciousness, by all accounts, present one of the great challenges and mysteries on earth, we need to be careful. If researchers were to begin the search by studying the hard problem of consciousness, arguably little progress would be made. Through theories like IIT and GNW, we move towards the wanted solutions. However, while

that is true, arguably there will be many dead ends, because the problem is greater than what was first assumed. Nagel does not end there. He also suggests that what if, due to the structure of our brains, we are actually unable to comprehend another conscious life form. While Nagel illustrates his point by comparing human consciousness to bat consciousness, it is easy to make the comparison to machines. In fact, one could argue, that it is an even greater challenge. The problem at hand is one of the deepest questions in nature. Some might say that by trying to create conscious machines, we try to replicate the essence of life itself. This is of enormous interest to science, of course, so should be contemplated deeply in whichever way we chose to go about it. Finally, Nagel points out how some phenomena has what he calls an "objective character", in stark contrast to consciousness. Consciousness is inherently about the "subjective character" of experience. Replicating this in machines, rather than simulating it is, one could say, one of the largest challenges in neuroscience, AI and computer science. On the contrary, if one believes that consciousness itself is an illusion, there is no challenge present at all.

Chalmers, while basing his definition of "the hard problem of consciousness" on Nagels definition, still goes about the problem of consciousness a little bit differently. Chalmers suggests that not only do we need to find ways of defining and understanding the phenomenological properties of experience, he also goes on to argue that maybe in fact we need to revise our understanding of the natural world, too. If consciousness does not fit into the natural world, maybe that is where the problem lies. This introduces an almost unfathomable depth to the problem. Consciousness already be-

ing a highly interdisciplinary search- we might have to include the fields of physics, chemistry and biology to the list as well. This does not apply for "the easy problem of consciousness", though. Therefore, when discussing this, it should be pointed out that there are different approaches one could have to the easy and the hard problem of consciousness. Thus, we might arrive at the conclusion that either both or one of the current neuroscientific theories of consciousness may very well explain the easy problem of consciousness, while the hard problem of consciousness requires different measures and approaches. What is interesting to contemplate is whether or not one is dependent on the other. Does the easy problem of consciousness only emerge in the space of the hard problem of consciousness? Or is it the case that the easy problem executes almost all functions of our consciousness except for the part about having an experience of being X? Regardless of what the answers to these questions might be, Chalmers does an outstanding job at laying out the different classes of the metaphysics of consciousness. By providing these, he gives a foundation for how science can approach the different possibilities at hand. For all we know, a complete theory of consciousness may be thousands of years into the future, and may cross some of the different classes even. By trying to create machine consciousness, we are able to experiment with different premises for human nature. Consequently, another grand challenge will be to come up with a test, like the Turing Test, but for consciousness. How do we measure the level of consciousness in a system? When it concerns the hard problem of consciousness, how do we measure the level of "what it is like to be X"? We need to face questions like the following: Is consciousness binary (lights on or lights off) or is consciousness

a spectrum? Is consciousness unfolding or emerging differently in computational systems than in humans? Is it even possible, without a biological body, to create consciousness?

What we may conclude with is that consciousness might be an area of research, in which all fields of science need to cooperate. Because the challenge is big, it might require different approaches and research methodologies in order to get there. We see this already happening now, with science borrowing ideas and conceptual frameworks from philosophy and vice versa. Not that that is a new thing, it has been that way with fields of research throughout history. But it might just be the case that in order to meet the challenge of consciousness, that different fields need even closer cooperation than what is required with various other problems that are being researched.

4. Ethical considerations

4.1. Should we create conscious machines?

We have now discussed whether or not it is possible to create conscious machines. One fascinating question we have yet to ask is: If it is possible to create conscious machines, should we? And if we do, how should conscious machines be treated? What rights should they receive? When are machines considered to be "conscious enough" to get human rights? What does it mean to have machines that have a "sense of self", an experience of being that machine?

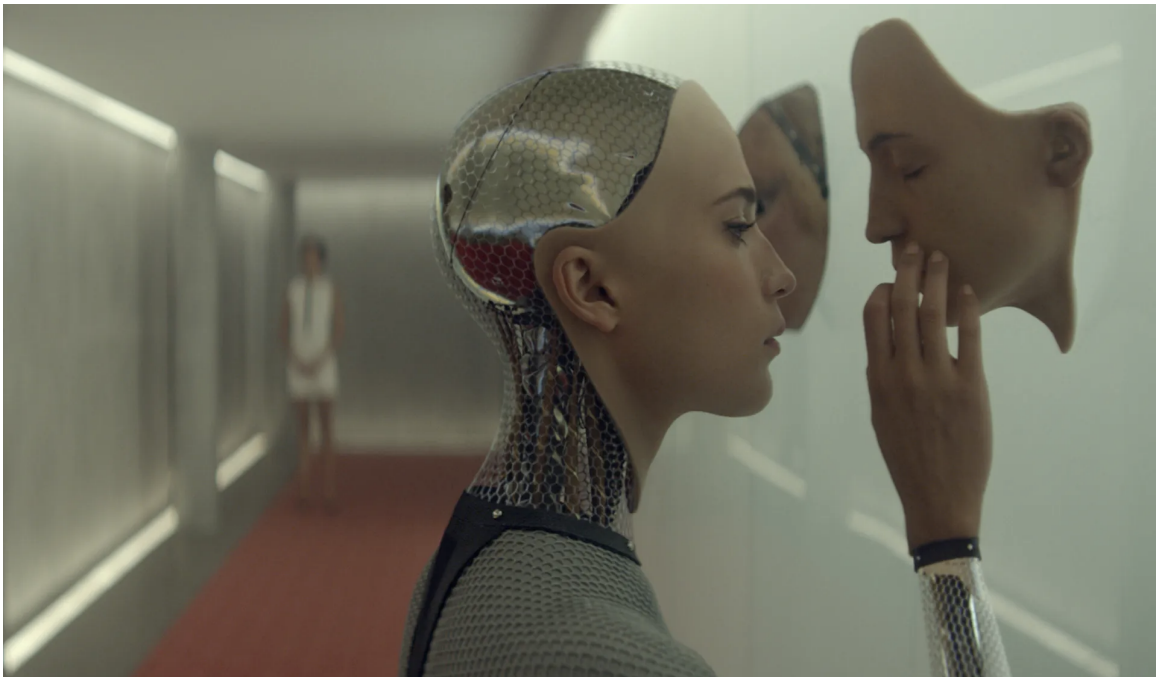

Figure 8: Ex Machina (2014)

There are really two questions here. The first one is a deep one, should we make conscious machines at all? One could argue that it is wrong to create systems that is able to suffer. If we could make these machines without consciousness, that would mean that they are inherently unable to suffer. Would that not be something to strive for? To not create more suffering in the world? One could say that on the other hand, if we are capable of creating conscious machines, we simply will. It is Murphy's law, *anything that can go wrong will go wrong*. Furthermore, another person may argue that it is immoral, or utterly wrong, to *not* make conscious machines. If

AIs will take over the world, would we not want that future world to have consciousness in it?

The second question is: Given that we will create conscious machines, how should they be treated? What rights should they have? What new ethical guidelines must be introduced, if any? Are we in danger of creating machines that are conscious but without us knowing? The idea of "self" is a complicated idea in philosophy, and thus here we will simply refer to a system having a sense of (an awareness of) its internal state and the external states that of its environment. It could also include that the system is able to discriminate itself from its environment. An entity's ability to feel pain and or/suffering is called the neurocentric criteria, and is often used as a reference point to guide which treatment should be given to different animals. How we morally regard (socially, ethically and legally) conscious machines, will be absolutely crucial if we create conscious machines [18]. If a conscious machine can suffer and feel pain, that they have a level of consciousness, that is something we as a society must take responsibility for.

5. Conclusion

We have now explored different theories of consciousness, both from the viewpoint of neuroscience and philosophy. By looking at consciousness at a closer proximity, we can divide matters of consciousness into easy and hard problems, or into the categorization of phenomenal consciousness and access consciousness. Through these divisions, we find clues as to how consciousness potentially can be implemented in machines. By looking at the many challenges with reducing consciousness into mathematical definitions, it was

contemplated whether or not machine consciousness is possible, and the possibility was explored that we may expand our conceptions of the natural world in order for consciousness to fit into it. While that may be the case, there are still many approaches to understand consciousness not through necessarily understanding the problem of *experience*, but through the realization that consciousness involves many subprocesses of which we can simulate, replicate and evolve. Consciousness may further be an emergent phenomena, that may require various scientific methods, approaches and explorations of different primitive rules to lay the right conditions for it to occur. Finally, we saw how many ethical implications conscious machines bring to our society and our world. Contemplating conscious machines is, as we have seen, a profound and meaningful pursuit, that allows us come face to face with some of the deepest questions about human nature.

References

[1] W. Commons, File:jørgen roed - an artist resting by the roadside - kms2063 - statens museum for kunst.jpg — wikimedia commons, the free media repository, 2020. URL: https://commons.wikimedia.org/w/index.php?title=File:J%C3%B8rgen_Roed_-_An_Artist_Resting_by_the_Roadside_-_KMS2063_-_Statens_Museum_for_Kunst.jpg&oldid=451140478, [Online; accessed 28-November-2021].

[2] A. Gianopoulos, Hubble gets galactic déjà vu, 2021. URL: https://www.nasa.gov/image-feature/goddard/2021/hubble-gets-galactic-deja-vu, last accessed 20 November 2021.

[3] D. J. Chalmers, Facing up to the problem of consciousness, Journal of consciousness studies 2 (1995) 200–219.

[4] M. O'Gieblyn, God, Human, Animal, Machine: Technology, Metaphor, and the Search for Meaning, Knopf Doubleday Publishing Group, 2021. URL: https://books.google.no/books?id=PxQLEAAAQBAJ.

[5] P. Ball, Neuroscience readies for a showdown over consciousness ideas, Quanta magazine 6 (2019).

[6] S. Maillé, M. Lynn, Reconciling current theories of consciousness, Journal of Neuroscience 40 (2020) 1994–1996.

[7] G. Tononi, An information integration theory of consciousness, BMC neuroscience 5 (2004) 1–22.

[8] R. Lemon, S. Edgley, Life without a cerebellum, Brain 133 (2010) 652–654.

[9] T. Nagel, What is it like to be a bat, Readings in philosophy of psychology 1 (1974) 159–168.

[10] D. J. Chalmers, Philosophy of mind: Classical and contemporary readings (2002).

[11] D. J. Chalmers, Consciousness and its place in nature, in: in Philosophy of Mind: Classical and Contemporary Readings, Citeseer, 2002.

[12] C. M. Signorelli, Can computers become conscious and overcome humans?, Frontiers in Robotics and AI 5 (2018) 121.

[13] B. Walker, The games that ai won and the progress they represent, 2020. URL: `https://towardsdatascience.com/the-games-that-ai-won-ff8fd4a71efc`, last accessed 10 November 2021.

[14] E. Hildt, Artificial intelligence: Does consciousness matter?, Frontiers in psychology 10 (2019) 1535.

[15] R. Chatila, E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, et al., Toward self-aware robots, Frontiers in Robotics and AI 5 (2018) 88.

[16] Y. Kinouchi, K. J. Mackin, A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot, Frontiers in Robotics and AI 5 (2018) 30.

[17] S. Franklin, T. Madl, S. Strain, U. Faghihi, D. Dong, S. Kugele, J. Snaider, P. Agrawal, S. Chen, A lida cognitive model tutorial, Biologically Inspired Cognitive Architectures 16 (2016) 105–130. URL: https://www.sciencedirect.com/science/article/pii/S2212683X16300196. doi:https://doi.org/10.1016/j.bica.2016.04.003.

[18] J. Giordano, Conscious machines? trajectories, possibilities, and neuroethical considerations, in: 2014 AAAI Fall Symposium Series, 2014.