The Connection Between Intelligence and Consciousness

The Challenges We Face

Thomas Pinella

4.20.2016
CSC 242W

Abstract

Our intuition tells us that there is a connection between intelligence and consciousness. We assume consciousness in other human beings and will generally grant it to other intellectually equipped, high-functioning mammals such as primates and dolphins. But we are more hesitant to attribute consciousness to lesser order beings, such as fruit flies and bacteria; relatively, they are lacking in the cognitive department. Although it is common for intuitions to be proven misleading, this one is not completely unfounded; as we will explore in the following pages, there is indeed a theoretical basis that supports the idea that intelligence and subjective experience are related at a fundamental level. Additionally, we will examine the intrinsic difficulties associated with the task of imbuing intelligence and, by extension, consciousness, into a machine.

Introduction

In his 1950 paper on computing intelligence -- a paper that ignited the field of artificial intelligence -- Alan Turing opens with a hypothetical game he called the “Imitation Game” (Turing 1950). The hypothetical game that originally involved

the interactions between three entities, a man, a woman, and an interrogator with a machine taking on the role of either the man or woman, has since transformed into the more popular “Turing Test.” This “Turing Test,” as it has come to be known, is a simplified version including one interrogator and one interrogatee. Through only conversation, it is the goal of the interrogator to correctly predict the identity of the interrogatee, whether it is human or machine, and it is the goal of the interrogatee to fool the interrogator into believing that is human. Should a machine succeed and trick the interrogator that it is human, we would consider the machine as having passed the “Turing Test” and, by the test’s definition, we would be forced to attribute it “intelligence.”

Although this test has been popularized over the years, especially with the Loebner Prize annual competition and its frequent appearances in Hollywood movies such as *The Imitation Game* and *Ex Machina*, the test is clearly not very scientific and leaves a very lacking definition of intelligence. But in some ways this is evidence of how difficult it is to properly formalize intelligence. All this being said, the “Turing Test” is a test for Strong AI, also known as Artificial General Intelligence (AGI). This intelligence is at the level of a human and is called “general” because of its ability to perform at a high level over a wide variety of tasks, like a human. This is opposed to Weak AI, which has a narrow non-general scope. In some ways then, the “Turing Test” is a reasonable test for this type of general intelligence, because what is more unpredictable or general than conversation?

The debate over whether or not it's possible or to build Strong AI, or AGI, has gone on for decades and will most likely continue for decades to come (Hawkins 2004). Those who argue that it is indeed possible often ask the logical question: "why should the substrate matter? If life can emerge from carbon, what should stop it from emerging from silicon?" After all, what is the brain but a complex organ that merely manipulates information? We may not fully understand it yet, but that's no reason as to why mimicking the functions of the brain in a computer should be impossible.

In order to find a more concrete connection between intelligence and consciousness beyond our mere intuition, we will first need to better understand and better formalize what it means to be intelligent. We will then introduce several theories on consciousness and examine how they relate to intelligence as will have formally defined it. Finally, we will take a look at the intrinsic difficulties associated with bringing the theories presented in this paper to life.

A Map to General Intelligence

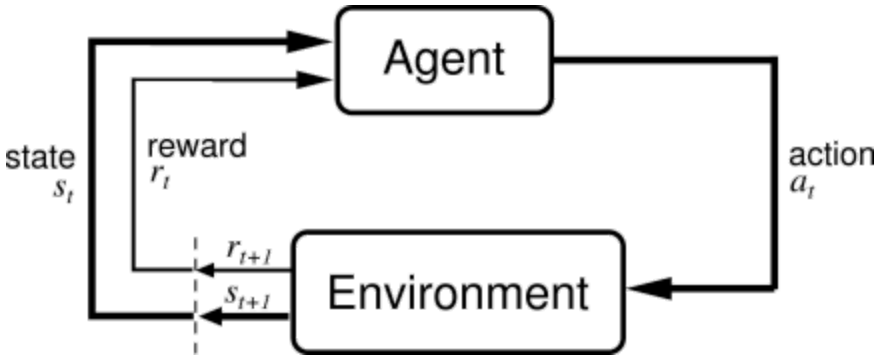
In 1964, Ray Solomonoff, a man considered to be one of the founders of algorithmic information theory (along with Claude Shannon), developed a mathematical method of universal inductive inference, referred to as Solomonoff

Induction. The concept, though uncomputable and impossible to implement on computers, was relatively straightforward: in order to predict what caused the current observation *x*, we test every possible cause (where each cause is viewed as an unhalting algorithm or a program *p*), and for those cases where the output of *p* matches *x*, the shorter *p*'s representation in code is, the more likely it is to be our hypothesis that explains *x* (Sunehag, P and Hutter, M 2011). This draws upon the idea of Occam's Razor, which states that the simplest explanation is generally the most likely to be the correct explanation.

From Solomonoff's theory of inductive inference came approximations of it, including one called AIXI created by Marcus Hutter. As Solomonoff Induction goes, AIXI tests all possible hypotheses (in actual implementation, it tests only a sample), but in addition to that, it uses reinforcement learning to work towards some goal by maximizing reward every iteration (Sunehag, P and Hutter, M 2011). This idea of using reinforcement learning in conjunction with Solomonoff Induction is shared by Schmidhuber's Godel Machine. Let's demystify reinforcement learning and see how it relates to artificial general intelligence.

Reinforcement Learning

This method of machine learning requires the distinction between an agent and its environment, and it involves how the two affect each other.



The important variables we need to keep track of are: states, rewards, and actions. The goal of the agent is to find the sequence of actions that maximizes its total expected sum of rewards. Schmidhuber took this model of reinforcement learning and used it for the purpose of building creative machines that are capable of coming up with novel and aesthetically pleasing constructions ranging from the arts to the sciences (Schmidhuber 2010). As we will see, this artificially manufactured creative power is directly linked to artificial general intelligence.

Computational Creativity

Without supervised learning, how is it possible to make use of reinforcement learning to build machines that possess some level of creativity? Schmidhuber approached this by attempting to formalize beauty. Essentially, the more something can be compressed, the more beauty it has. Beauty, by this definition then, is directly related to the Kolmogorov complexity (the length of the most compressed version of some data) of the object in question. In fact, this brings us

back to Solomonoff Induction: the shortest explanations are the most probable predictions. The deep ties between compression and prediction are commonly understood in the field of algorithmic information theory (Franz 2015). Therefore, that which is most predictable is also the most compressed, and, by extension, the most beautiful. In Schmidhuber's construction of an agent that is creative, however, he does not set beauty as the reward to maximize. Instead, his reinforcement learning algorithm maximizes the first derivative of beauty, which he refers to as "aesthetic pleasure." This can be visualized in the following equation where *O* denotes the observer, or agent, at time *t*, *D* is the observed phenomenon, *B(D, O(t))* is the compressed observation or beauty as we defined it previously, and *I(D, O(t))* is the "aesthetic pleasure" garnered by the agent as it observes the change in beauty:

$$I(D,\ O(t)) = \frac{\delta B(D, O(t))}{\delta t}$$

By maximizing the change in beauty, seen above as *I(D, O(t)),* the agent discovers or creates things that are not only beautiful, but are also novel. Because of this, it does not get stuck creating the same thing, even if it's beautiful, over and over again; it gets "bored" and seeks to make new things.

An important detail: in order to maximize the change in beauty, the agent does this by modifying and improving its compressor, making it ever more efficient. The general intelligence present in this model grows clearer in that we see that

the agent is acting like a scientist. What does a scientist do? A scientist strives to distill the world he or she observes into simple rules; the scientist, like our agent, is compressing. And there we have our definition of general intelligence: intelligence is the process of observing the world and compressing it into a simpler model capable of accurate prediction.

The Question of Consciousness

Although the concept of general intelligence can be a difficult one to grapple with and define, it is easy in comparison to understanding the nature of consciousness. In consciousness research, a small but growing field, there are commonly understood to be two types of problems to be dealt with: the “easy” problems of consciousness and the “hard” problems of consciousness (Chalmers 1997).

Easy Problems of Consciousness

First, the easy: these are problems that neuroscience has been in the process of solving for decades, and they have been successful so far. These problems ask for the correlation between observable behavior and activity in the brain; essentially, they ask how brain mechanisms perform functions. For this reason,

the scientific method has worked for neuroscientists interested in finding exactly what brain activities are responsible for certain behavior. Questions that fall into this category include how entities categorize and react to environmental stimuli, the difference between awake and asleep, and the ability to control behavior (Chalmers 1997).

Hard Problems of Consciousness

The “hard” problems of consciousness cannot be solved using classical objective scientific methods. This domain of problems has to do with the *why*’s and *how*’s of subjective experience. The term “qualia” is often used to describe our subjective experience of the world around us. The “hard” problems of consciousness seek to understand things such as why we see the color red as red, and how this comes to be. Because they have to do with the subjective, not the objective, David Chalmers, the philosopher who classified consciousness-related questions into these two distinct domains, believes that it is fundamentally impossible to arrive at a satisfactory answer to the “hard” problems of consciousness through current physics and the scientific method. Nevertheless, over the years people from all different fields, from neuroscience to philosophy to computer science, have taken a stab at this seemingly impossible task.

Consciousness as Fundamental: Panpsychism

In an attempt to formulate his own theory for consciousness, David Chalmers circumvented the issue of physics being unable to solve the mystery by proposing his own law of physics: that consciousness is a fundamental principle of matter. His reasoning follows the idea that if something cannot be explained or broken down any further, it is worth making fundamental. He cites electricity as an example of this. We understand the effects of electricity and numerous properties of the phenomenon, but at a deep enough level, we just need to accept its clear existence. So too with consciousness, Chalmers argues. Such a view that consciousness is intrinsic in nature is considered to be panpsychism.

Even if this idea of panpsychism were to be held as true, this still would not explain how consciousness works in our own minds. While some, such as Chalmers, ponder very raw theories of consciousness as it exists fundamentally in nature, others take a more cognitive-based approach, looking at consciousness in the context of the human mind.

The Theater of Consciousness

An analogy of consciousness introduced by neurobiologist Bernard Baars, this

take on consciousness, although it does not supply answers to any of the fundamental questions, provides an easy to understand idea of how consciousness operates in the context of our brain. Baars compares the mind to a theater, complete with a stage, actors, director, and an audience. He relates the spotlight on the stage as your conscious awareness and the various actors as your senses, thoughts, and ideas. He emphasizes that at any given moment, your consciousness is very narrow and you are aware of only one thing. For example, as you read these words, your eyes saccade from word to word, phrase to phrase, but at any given moment in time, only a single word or two are present, or illuminated, by the spotlight that is your conscious mind. Additionally, to continue Baars theater analogy to consciousness, the audience makes up the vast unconscious part of one's mind; it includes the deep beliefs of the individual as well as all memories. The unconscious guides the conscious mind and informs it. When you see a friend, for example, your unconscious retrieves the memory of him or her and brings it to your consciousness, allowing you to recognize him or her. This is also where Baar's theater metaphor breaks down a bit, since the audience in a theater generally doesn't guide the actions of the actors on stage. Finally, the director would be analogous to the mind's sense of self. It is the director that runs the show and calls the shots (Baars 1997).

While Baars clearly points out that the spotlight in his analogy is the zone of consciousness, he never explains how that light shines; he does not explain where consciousness originates from.

Self-Model Theory of Consciousness

Yale computer science professor Drew McDermott proposed the theory that a self-model is necessary for an entity to experience consciousness. The evolutionary purpose of the self-model would be that it informs the entity on how to interact with the world. As the theory goes, the entity would hold an entire world model and within that world model would naturally exist a representation of the self: the self-model. In fact, within that self-model would exist another world model which would house a secondary, but one layer extra abstraction of a self-model , and we find ourselves in an infinite recursion. McDermott makes it clear that the self-model theory of consciousness is not merely consciousness of the self, but an explanation as to how consciousness is experienced in the first place. As McDermott states: “Phenomenal consciousness is not part of the mechanism of perception, but part of the mechanism of introspection about perception.” It is this “introspection” step -- a step that requires a self-model -- that McDermott believes leads to consciousness (McDermott 2007).

The theories outlined above and the vast majority of theories about consciousness reside in the world of subjective descriptions that can be hard to truly formalize, and therefore hard or impossible to use to make any useful predictions (Metzinger 2003). One exception to this is a theory that is

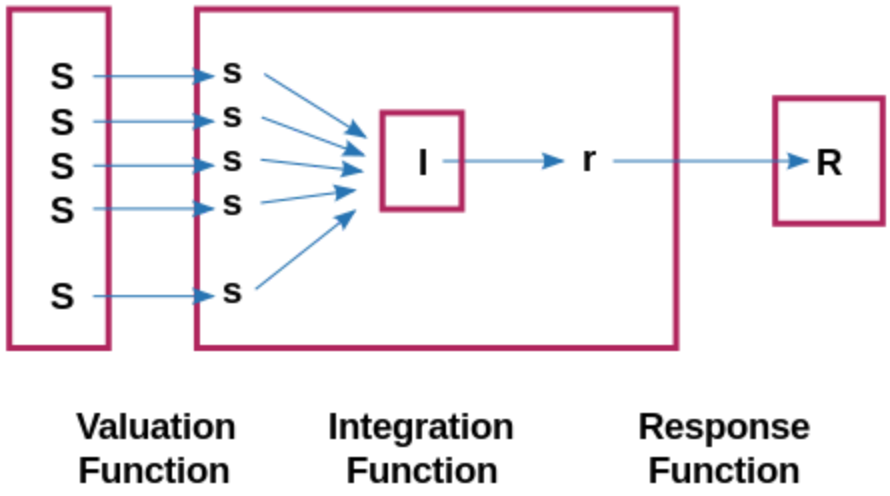
mathematically grounded, taking the subjective and describing it in the universal, objective language of mathematics. It is called the Integrated Information Theory of Consciousness (Tonini 2008).

Integrated Information Theory of Consciousness

One of the most highly debated, but also highly regarded theories involving consciousness, is the Integrated Information Theory of Consciousness (IIT), developed by Giulio Tononi, a neuroscientist from the University of Wisconsin. The theory states that consciousness arises when systems are able to take in information and integrate, or unify them, such that the result is more than the sum of its parts. In essence, Tonini argues that the experience of information, specifically integrated information, is what gives rise to consciousness (Tononi 2008).

The example Tonini gives in his manifesto on IIT is being in an empty, dark room alongside a simple light sensing photodiode. When the lights are turned on, you would of course experience the light, but the photodiode would also experience it. Your experience versus the experience of the photodiode would be fundamentally different, however. When you perceive just the simple lightness, you are discriminating this state of experience from countless other possible states you could be experiencing, and since information is the reduction of

uncertainty, you are therefore experiencing an abounding amount of information. The photodiode, on the other hand has only two possible experiences: lightness and darkness. In fact, it would not even know it is lightness or darkness, just as it would not know if it is sensing hotness or coldness. Therefore, when it perceives the lightness, we can use the entropy function and take the $log_2 2$ to get that it is experiencing exactly one bit of information (we take the log of two because it can only be in two possible states that we are considering to be equally likely). The amount of information something is able to perceive or the richness of its consciousness, according to this theory, can actually be calculated and is called Φ (Tonini 2008).



By Traced by User:Stannered - en:Image:Information-integration.png, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1831275

What it boils down to, under this theory, is integration. A video camera, which is taking in a high amount of raw data and is essentially an array of photodiodes, is

not any more conscious than a single photodiode for the reason that it does not integrate any of the information it sees. There is no unification taking place. For example, if you were to record a car driving down the road, you could somewhat easily remove the car from the camera's memory by going in and erasing all the pixels that caught the car on film. However, if a human being, whom we are assuming to be a conscious entity, were to see a car driving down the road, and you wanted to erase the car from his or her memory, you would find that to be a much more difficult, if not impossible, task. The image of the car has already been integrated with the subject's memories all across the brain, and there is no obvious way of being able to pick out the parts of the car without affecting anything else (Maguire, P., et al. 2014).

Although it's not panpsychism, the implications of IIT are startling. It means that any system, not necessarily a biological system, can be conscious, and its Φ can show us exactly how conscious it is.

Calculating Φ

Given a system where nodes can be in one of two states and certain nodes have a causal relationship with other nodes they are connected to, we first need to establish two probability distributions: the potential repertoire (expressed as $p(X_0(maxH)))$) and the actual repertoire (expressed as $p(X_0(mech, x_1))$). The

potential repertoire consists of every possible combination of node states across the entire system and assumes an equal probability of each. For example, if there are two nodes, with each either ON or OFF, there would be a total of 2^2 or four possible states. The potential repertoire probability distribution, therefore, would be $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The actual repertoire would depend on the predetermined causal relationship between the nodes. Continuing with our previous example, if it were assumed that node1 had to always be ON, then that limits our total possible states to just two total, making our actual repertoire probability distribution $(0, 0, \frac{1}{2}, \frac{1}{2})$. We then find the relative entropy between these two probability distributions, giving us a measure on the difference between the two distributions that would come out to be zero if they were identical. This results in what IIT calls effective information (ei). Thus, the full equation for ei is as follows:

$$ei(X(mech,\ x_1) = H[p(X_0(mech,\ x_1))||p(X_0(maxH)))]$$

In order to calculate Φ, however, we must look at how much of the information is integrated information. To do this, we must measure how much information is generated by each part of the system and compare it to the information generated by the system as a whole (Tonini 2008).

To do this, we once again use the relative entropy function in order to find the difference between the entire system's actual repertoire and the product of each part's actual repertoire. In full, it looks like this:

$$\Phi(X(mech,\ x_1)) = H[p(X_0(mech,\ x_1))||\Pi p({}^{k}M_0(mech,\ \mu_1))]\ for\ {}^{k}M_0 \in MIP$$

Given any system, we can use the above equation to measure its Φ, and thereby, according to IIT, determine how conscious it is.

Criticism of IIT

The most noteworthy criticism of IIT comes from the blog of theoretical computer scientist Scott Aaronson. In 2014, he published what he believed to be a counterexample that proved IIT failed to provide a reasonable theory for consciousness (Aaronson). First, he makes it clear that IIT is not even claiming to have solved the "hard" problem of consciousness; nowhere does it specify exactly how or why we experience qualia the way we do. Instead, Aaronson suggests that it attempted to solve what he called the "pretty hard" problem of consciousness: namely the question of what types of physical systems give rise to consciousness. This is, after all, exactly what IIT does with Φ; it informs us how conscious a physical system is.

But Aaronson continues. He claims that IIT fails to do this successfully. He provides an example showing that a simple *n x n* network of XOR gates would generate a Φ of $\sqrt{n}$. Therefore, given an arbitrarily large *n*, you could theoretically have a vast, but extraordinarily simple, array of connected XOR

gates and call it more conscious than a human being. The fact that this is possible given the math for Φ, completely discredits the theory for Aaronson as he finds it absurd that a mere network of logic gates could be considered more conscious than a human if large enough. Tonini responded to Aaronson's blog post stating that given the theory, which he still stands by, seemingly unintuitive constructs can be indeed conscious. The issue here, it would be appear, is the lack of a mutual definition of consciousness; Aaronson seems to be expecting something very much different than what Tonini is proposing.

Different Consciousnesses

In Aaronson's critique of IIT, Aaronson wonders what relevancy Φ has, given some of its very backwards and unintuitive predictions. The rift in viewpoints between Tonini and Aaronson can be better seen in their definitions of consciousness. Tonini views consciousness closer to the way Chalmers does: as something more fundamental to nature, whereas Aaronson is looking for a theory of consciousness that is more cognitive-based, something closer to what Baars presents in his theater analogy of consciousness or what McDermott writes about when he speaks of an entity using its self-model to inform decisions.

From the theories we have seen so far, there are three definitions or types of consciousnesses we can consider. First is the very fundamental, panpsychic

“proto-consciousness” that Chalmers describes; it is intrinsic to every quark and is incomparable to our own experience of consciousness. Next, is Tonini’s version of consciousness which is also a form of “proto-consciousness” in that it can arise in unlikely places, such as Aaronson’s XOR network, and its subjective experience would be completely unlike our own (Cerullo 2015). Finally, the third type of consciousness is the one we are most familiar with; it is the one neuroscientists seek to understand and is the one Baars and McDermott wrote about. It is also what Aaronson was looking for in Tonini’s theory. Rather than a raw “proto-consciousness,” it is a “cognitive-consciousness” that is capable of not only raw experience, but is also associated with a self or mind. This third type of “cognitive-consciousness” can be further subdivided into two groups: access and phenomenal consciousness where the former has the ability to represent content with thoughts, beliefs, memories, etc. and the latter is associated only with sensations (Block 1995).

With this expanded perspective on consciousnesses, we now see that while Aaronson was attacking IIT as a failed theory attempting to solve the “pretty hard” problem of cognitive-consciousness, Tonini was defending it as a successes in answering the problem for a simpler “proto-consciousness” (Cerullo 2015).

Although both parties may have been correct, this starts to question the relevancy of IIT and Φ. What use is measuring “proto-consciousness?”

IIT and Intelligence

It turns out that the integration a system must perform on the incoming information is directly comparable to the compressing we needed to perform in order to achieve the definition of general intelligence. Herein, we see the connection between consciousness and intelligence: both require the ability to compress their observations of the world: we called it integration in the study of consciousness and it opens the doors for prediction in the realm of intelligence (Maguire, et al. 2014).

While a system that possesses general intelligence will necessarily contain a high Φ, the same cannot be said in reverse order; a system that possesses a high Φ is not necessarily intelligent. And we saw this in Aaronson's *n x n* grid of XOR gates. Assuming that Φ is a measure on the level of "proto-consciousness" of a system, then it follows that "proto-consciousness" is the natural byproduct of an intelligent system, but not all "proto-conscious" systems are necessarily intelligent.

Self-Models and Intelligence

In order for a compressor to be effective, the agent performing the actions will generally hold a representation of the agent itself within itself. This way, the agent is capable of saving a total history of knowledge that it has experienced. By

referring to a model of the self, it is also fulfilling McDermott's definition of consciousness, which is a "cognitive-consciousness."

Therefore, as we can see, an entity capable of general intelligence not only naturally has "proto-consciousness," as evident in a high Φ, but it would also have a more familiar "cognitive-consciousness," the type of consciousness that more resembles our own experience.

It would appear that we have a formula to follow if we want to create intelligent conscious machines: build a reinforcement learning program that maximizes its reward such that it is improving and modifying its compressor algorithm making it capable of accurate prediction and given that the agent contains a model of itself in order to improve its compressor. There are, however, a couple problems.

Problem I: Mysterianism

Phil Maguire et al. investigate the computability of a general compressor that is capable of such high integration and prove that such a compressor is not computable (2014). It should be possible to take the integrated memories of a person and reverse engineer them to deduct what the raw input information originally was. Maguire et al., however, show that it cannot be computationally modelled. The reason for this is somewhat odd to fathom; it's not due to some

magic that occurs in the brain, but rather to our own inability to formalize what is happening. It is the same reason a dog, for example, is simply unable to comprehend multivariable calculus. We are just not properly mentally equipped (Kriegel).

But say we are somehow able to discover the optimal compression algorithm capable of complex integration, or it is given to us by some alien species (after all, just because we cannot comprehend it does not mean another hypothetically more cognitively advanced species cannot). In that case it is *still* theoretically possible, however unlikely, to build an intelligent and conscious entity. But, one final problem remains. And this one has no clear workaround.

Problem II: The Fabric of Reality

This entire time we have been assuming a purely causal, deterministic universe. Unfortunately for us, and our theoretical conscious machine, that is simply not the universe we live in. For if we did live in such a universe, every single detail about the current state of the universe could be traced back as the effects of the initial conditions of when the universe was born. But due to the discovery of quantum mechanics, we know this not to be true. With this more scientifically accurate perspective, we observe that there is indeed such a thing as true randomness, and it becomes increasingly difficult to view consciousness as an

emergent epiphenomenon. There are generally three interpretations of quantum mechanics, namely they are Everett's Interpretation, Wigner's Interpretation, and Bohm's Interpretation (Chalmers 1996). Everett's Interpretation describes a "many worlds" scenario where for each event, a separate universe splits off and all possible iterations occur over all universes. Wigner's Interpretation places the emphasis on consciousness itself, claiming that even macroscopic objects are in superpositions when gone unobserved. Finally, Bohm's Interpretation, though the least radical, is mathematically the clunkiest and posits that hidden variables are to blame for quantum behavior. Einstein, who was vehemently against quantum mechanics, held similar views.

Regardless of how we interpret the quantum phenomenon we observe, however, it forces us to reframe how we view consciousness, especially Wigner's Interpretation, which places consciousness as the most fundamental thing, more so than matter.

Conclusion

Although the only conscious entity you can be sure is conscious is yourself, we naturally ascribe consciousness other human beings by analogy and we intuitively attribute it to the animals most like us -- mammals such as primates and dolphins. But as you go down the ladder of species, we become less and less

likely to assume consciousness, to the point that most would assume it certainly does not exist, like in the case of an amoeba. As we have seen, there is indeed a correlation not only between general intelligence and "proto-consciousness" (although we saw that "proto-consciousness," as defined per IIT, is not necessarily intelligent), but also between general intelligence and the more familiar "cognitive-consciousness." However, even though most believe it to be theoretically possible to produce intelligent, and by extension, conscious, machines, we run into a couple practical, but unavoidable, problems. Namely, the one of mysterianism where we simply cannot formalize what we are observing, and the issue that we do not truly have a firm grasp on what factors may be at play in the unravelling of the universe and consciousness.

References

A. M. Turing (1950) *Computing Machinery and Intelligence. Mind 49: 433-460.*

Baars, B. J. (1997). In the theater of consciousness: The workspace of the mind. New York: Oxford University Press.

Baars, B. J. (1988). A cognitive theory of consciousness. New York: Guilford Press.

Block N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences 18: 227–287.

Cerullo, M. A. (2015). The Problem with Phi: A Critique of Integrated Information Theory. *PLoS Computational Biology*, *11*(9), e1004286. http://doi.org/10.1371/journal.pcbi.1004286

Chalmers, D.J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies*

Chalmers, David John. *The Conscious Mind: In Search of a Theory of Conscious Experience*. New York: Oxford UP, 1996. Print.

Franz, A. (2015). Artificial general intelligence through recursive data compression and grounded reasoning: a position paper.

Hawkins, Jeff, and Sandra Blakeslee. *On Intelligence*. New York: Times, 2004. Print.

Kriegel, U. Mysterianism. For the *Oxford Companion to Consciousness*

Maguire, P., et al. (2014). Is Consciousness Computable? Quantifying Integrated Information Using Algorithmic Information Theory.

Metzinger, Thomas. *Being No One: The Self-model Theory of Subjectivity*. Cambridge, MA: MIT, 2003. Print.

McDermott, D. (2007). Artificial Intelligence and Consciousness. *The Cambridge Handbook of Consciousness.* Cambridge University Press

Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation. *IEEE Transactions on Autonomous Mental Development, Vol. 2, No. 3*

Sunehag, P and Hutter, M. (2011). Principles of Solomonoff Induction and AIXI. *Research School of Computer Science, Australian National University Canberra, ACT, 0200, Australia*

Tononi, G. (2008). Consciousness as Integrated Information: a Provisional Manifesto. *Biol. Bull. 215: 216 –242*

Tononi G, Koch C. 2015 Consciousness: here, there and everywhere? Phil. Trans. R. Soc. B 370: 20140167.

"A Brief Introduction to Reinforcement Learning." *A Brief Introduction to Reinforcement Learning*. N.p., n.d. Web. 28 Mar. 2016.

"Shtetl-Optimized." *ShtetlOptimized RSS*. N.p., n.d. Web. 20 Apr. 2016.

"9 Temporal-Difference Learning." *9 Temporal-Difference Learning*. N.p., n.d. Web. 28 Mar. 2016.