Reber, Arthur S. 1997. "Caterpillars and consciousness." *Philosophical Psychology* 10, no. 4: 437. *Academic Search Complete*, EBSCO*host* (accessed March 17, 2016).

**Abstract:** The dominant position in the field of artificial intelligence (AI) is computationalism where the operative principle is that cognition in general and consciousness in particular can be captured by identification of the proper set of computations. This position has been attacked from several angles, most effectively, in my opinion, by John Searle in his now famous Chinese Room thought experiment. I critique this Searlean perspective on the grounds that, while it is probably correct in its essentials, it does not go far enough. Quite simply, it runs afoul of the problem of emergentism. The proffered solution to this problem is that consciousness (or very rudimentary forms of it) needs to be viewed as an inherent property of organic form. While this recasting of the problem solves the emergentist dilemma it opens up a number of other issues. However, the new problems, unlike the old, appear in principle to be amenable to scientific analysis.

**Annotation:** This source goes over the anthromorphism of machines using human traits such as consciousness. By being able to make connections using with our own behavior, it is easier to analyze difficult concepts. In terms of composure, this source uses credible papers in order to establish authority and focuses on explaining core concepts such as emergentism and the difference between strong and weak AI. In terms of weaknesses, this paper also spends a lot of time paraphrasing and does not contain much physical data, also his reasoning may seem to be highly opinionated at times. A strength of this paper is the use of colloquial language in order to clarify his reasoning. This is highlighted through this statement, “But since I think that Searle is basically correct, that the strong AI gang is misguided, we have a problem and a serious one for one thing is pretty clear: we don't have a clue as to how complex mental life can emerge from a brain.” This paper analyzes important documents within the area of artificial intelligence and explains a solution to emergentist dilemma is to focus primarily on explaining complex systems but not being stopped by small regions which cannot be hypothesized or miraculously explained. Within my paper, I will use it to highlight important aspects of emergentism as it defines the new or unexpected outcomes and use it to explain how consciousness may arise in machines not due to our understanding of consciousness, but through unexpected changes.

Piccinini, Gualtiero. 2010. "The Resilience of Computationalism." *Philosophy Of Science* 77, no. 5: 852-861. *Academic Search Complete*, EBSCO*host* (accessed March 17, 2016).

**Abstract:** Computationalism—the view that cognition is computation—has always been controversial. It faces two types of objection. According to insufficiency objections, computation is insufficient for some cognitive phenomenon X. According to objections from neural realization, cognitive processes are realized by neural processes, but neural processes have feature Y, and having Y is incompatible with being (or realizing) computations. In this article, I explain why computationalism has survived these objections. To adjudicate the dispute between computationalism and its foes, I will conclude that we need a better account of computation.

**Annotation:** This source goes over how computationalism and Theory of Mind overlap and how we can define these nuances in order to reach the most correct definition. In terms of composition, the author attempts to make connections between cognition and computation and also raises prominent counter-arguments and explains their invalidity. A common counter-argument raised is the objection from neural realization. In this argument, computationalism and neural process cannot be aligned because of the nonelectrical neurotransmitters present in the brain. The author explains how this is false because computations are realizable through an indefinite number of physical substrates, so this argument is not a true limitation. The strengths of this paper can be seen through the use of counter-arguments as it reduces bias and also strengthens the author’s argument by researching and analyzing differing viewpoints. The weaknesses of this paper are the in the section where the author explains how to test computationalism but does not give a clear answer of what occurs during his tests. These tests are more so rationalization that actual testing. I plan to use this source in my paper to compare and contrast computationalism and Theory of Mind and also provide skepticism on the importance of declaration with respect to emergentism.

SEBASTIÁN, M. Á. (2016). Consciousness and Theory of Mind: a Common Theory?. Theoria, 31(1), 73-89. doi:10.1387/theoria.14091

**ABSTRACT:** Many have argued that the difference between phenomenally conscious states and other kind of states lies in the implicit self-awareness that conscious states have. Higher-Order-Representationalist (HOR) theories attempt to explain such self-awareness by means of higher-order representation. Consciousness depends on our capacity to represent our own mental states: our Theory of Mind. It is generally agreed that such an ability can be decomposed into another two: mindreading and metacognition. I will argue that consciousness cannot depend on mindreading. The tenability of HOR theories depends, therefore, on the relation between mindreading and metacognition. I analyze several views on such a relation and argue that none of them seem to be a plausible option for HOR theories. Keywords: Consciousness, Self-awareness, Higher-Order Theories, Theory of Mind, Mindreading, Metacognition

**Annotation:** This paper discusses the differences in properties of mental states and how it factors into our definition of consciousness. This paper maintains its relevance as it brings to light issues with Theory of Mind as it debates phenomenal consciousness must be reached by Theory of Mind. Within this paper, the author defines consciousness as the ability for humans to be able to identify their own mental states and also differentiate their own mental states from others. The author provides the definition of consciousness as mindreading and metacognition. Being able to predict what another animal is thinking was traditionally a defense mechanism that humans acquired and differs from our definition of animal consciousness. By understanding that you exist and also understanding that other beings are conscious, by definition, you can be labeled as conscious. This article supports my research as it displays one example of consciousness within the humanities field and can be contrasted with our definition of consciousness in AI.

Dournaee, Blake H. 2010. "Comments on “The Replication of the Hard Problem of Consciousness in AI and Bio-AI”." Minds & Machines 20, no. 2: 303-309. Academic Search Complete, EBSCOhost (accessed March 17, 2016).

**Abstract:** In their joint paper entitled ‘‘The Replication of the Hard Problem of Consciousness in AI and BIO-AI’’ (Boltuc et al. Replication of the hard problem of conscious in AI and Bio-AI: An early conceptual framework 2008), Nicholas and Piotr Boltuc suggest that machines could be equipped with phenomenal consciousness, which is subjective consciousness that satisfies Chalmer’s hard problem (We will abbreviate the hard problem of consciousness as ‘‘H-consciousness’’). The claim is that if we knew the inner workings of phenomenal consciousness and could understand its’ precise operation, we could instantiate such consciousness in a machine. This claim, called the extra-strong AI thesis, is an important claim because if true it would demystify the privileged access problem of first-person consciousness and cast it as an empirical problem of science and not a fundamental question of philosophy. A core assumption of the extra-strong AI thesis is that there is no logical argument that precludes the implementation of H-consciousness in an organic or in-organic machine provided we understand its algorithm. Another way of framing this conclusion is that there is nothing special about H-consciousness as compared to any other process. That is, in the same way that we do not preclude a machine from implementing photosynthesis, we also do not preclude a machine from implementing H-consciousness. While one may be more difficult in practice, it is a problem of science and engineering, and no longer a philosophical question. I propose that Boltuc’s conclusion, while plausible and convincing, comes at a very high price; the argument given for his conclusion does not exclude any conceivable process from machine implementation. In short, if we make some assumptions about the equivalence of a rough notion of algorithm and then tie this to human understanding, all logical preconditions vanish and the argument grants that any process can be implemented in a machine. The purpose of this paper is to comment on the argument for his conclusion and offer additional properties of H-consciousness that can be used to make the conclusion falsifiable through scientific investigation rather than relying on the limits of human understanding.

**Annotation:** In order to further understand the possibilities of consciousness and machine implementation, we must first be able to provide an algorithm since they can always be applied to machines. Here within this paper, we are able to see that the author focuses primarily on analyzing Boltuc’s conclusions and seeing how plausible they may be. Essentially, he understands that we have to increase specificity of our overall definition of consciousness in order to be able to apply it correctly to machines. The strengths of this paper are its commentary as it provides insight on how to analyze a paper and also reach your own conclusions from similar sources. The weaknesses are the lack of applicable testing methods and also the fact that it limits its sources to Boltuc and does not provide many counter-arguments. I will use this as a source to provide contrast between our understanding of the application of algorithms and how we define general consciousness and see if the two have any overlapping areas which can support machine consciousness.

Perri. 2001. "Ethics, regulation and the new artificial intelligence, part II: autonomy and liability." *Information, Communication & Society* 4, no. 3: 406-434. *Academic Search Complete*, EBSCO*host* (accessed March 17, 2016).

**Annotation:** This is the second article in a two-part series on the social, ethical and public policy implications of the new artificial intelligence (AI). The first article briefly presented a neo-Durkheimian understanding of the social fears projected onto AI, before arguing that the common and enduring myth of an AI takeover arising from the autonomous decision-making capability of AI systems, most recently resurrected by Professor Kevin Warwick, is misplaced. That article went on to argue that, nevertheless, some genuine and practical issues in the accountability of AI systems that must be addressed. This second article, drawing further on the neo-Durkheimian theory, sets out a more detailed understanding of what it is for a system to be autonomous enough in its decision making to blur the boundary between tool and agent. The importance of this is that this blurring of categories is often the basis, the first article argued, of social fears.

**Annotation:** This source raises concerns about the use of autonomous machines and questions the ethical nature of using AI for public policy and also raises questions about how we would ensure liability. This connects to current development projects on using AI for autonomous drones. The same questions of liability are also present as how can we ensure that there are no systems in place which are prone to error in such a situation. This is relevant as the AI is a growing area of CS as its applications continue to grow to be used for more and more processes. This paper uses brief sources to explain the several types of autonomy and also compares and contrasts each versions. In terms of strengths, it provides great insights into definitions and utilizes a lot of connection making in order to clarify concepts. In terms of weaknesses, most of this paper relies on opinions/skepticism and can be clearly argued against. I will use this paper as a way to raise questions these ethical concerns may validate the theory of machine consciousness.

Question: How universal is our understanding of consciousness with respect to machines?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Source | Summary | Definition of Consciousness | Response v.s conscious thought | Testing methods | Diverging theories | Ethical Concerns |  |
| Caterpillars and Consciousness | This source goes over the anthromorphism of machines using human traits such as consciousness. By being able to make connections using with our own behavior, it is easier to analyze difficult concepts. | Author relates consciousness to simple traits and states that if we are able to find connections between human behavior, we can define it as consciousness |  | No testing, uses anecdotal evidence at times, but focuses on providing definitions over testing. | By anthromorphising machine structures, we are able to better define general consciousness. |  |  |
| The Resilience of Computationalism | This source goes over how computationalism and Theory of Mind overlap and how we can define these nuances in order to reach the most correct definition. | Combines consciousness and computationalism into one theory as they are the same. Both require similar elements as consciousness requires computation. |  |  | Computationalism is Concsiousness |  |  |
| Consciousness and Theory of Mind: a Common Theory? | This paper discusses the differences in properties of mental states and how it factors into our definition of consciousness. This paper maintains its relevance as it brings to light issues with Theory of Mind as it debates phenomenal consciousness must be reached by Theory of Mind. |  |  |  |  |  |  |
| Comments on “The Replication of the Hard Problem of Consciousness in AI and Bio-AI | **:** In order to further understand the possibilities of consciousness and machine implementation, we must first be able to provide an algorithm since they can always be applied to machines. Here within this paper, we are able to see that the author focuses primarily on analyzing Boltuc’s conclusions and seeing how plausible they may be. |  |  | No testing methods, focuses primarily on analyzing the research of Boltuc. |  |  |  |
| Ethics, regulation and the new artificial intelligence, part II: autonomy and liability | This source raises concerns about the use of autonomous machines and questions the ethical nature of using AI for public policy and also raises questions about how we would ensure liability. This connects to current development projects on using AI for autonomous drones. |  |  |  |  | Should we allow AI to be provide influence in our decision making? Is AI advanced enough to avoid errors and differentiate between right and wrong? |  |