- I believe that it believes that I believe - Dennett's Answer to the Objection of Original Intentionality in Artifacts

Philosophy & Artificial Intelligence
Janosch Haber
University of Amsterdam
10400192
janoschhaber@gmail.com

March 12, 2018

This paper is intended to defend the idea of *intentionality* in *digital computers* through arguments from contemporary philosopher Daniel Dennett's *Intentional Systems Theory.*

1 Introduction

Many critics of AI argue that intentionality in computers - or any other artifact for that matter - can never be more than *derivative.* With the words of John Haugeland, their "tokens only have meaning because we give it to them" and consequently, "they only mean what we say that they do" (Haugeland, 1981, p. 15). Contemporary philosopher Daniel Dennett (2011, p. 8) however claims that "there is no principled (theoretically motivated) way to distinguish 'original' intentionality from 'derived' intentionality." On the basis of this idea, he developed a three-stage model to explain the assignment of intentionality and refute the objection of derived intentionality in artifacts.
In this paper, we analyze Dennett's model and answer the question

> *How does Dennett's elaborated model of the intentional stance answer Haugeland's objection that intentionality in artifacts cannot be original?*

Approaching this question will be guided by the Hypothesis that

> *Dennett's Intentional Stance Model renders a valid approach to refute the derived intentionality argument and therefore re-opens the theoretical possibility of intentionality in computers.*

In the first section, we will have a short look at the background of Haugeland's claim that computers cannot have meaning, which he formulated when working on *Automatic Semantic Engines.* Next we give a detailed overview of Dennett's *Intentional Systems Theory* and especially his idea of intentionality assignment through the *Intentional Stance.* We then analyze how this model answers Haugeland's objection and consider some possible caveats and pitfalls of Dennett's theory. We conclude that Dennett's theory is sufficient to at least theoretically allow for the possibility of intentionality in computer systems and therefore makes it possible for AI researchers to not lose their faith just yet.

2 Automatic Semantic Engines

In his introduction to 'Mind Design,' philosopher John Haugeland analyzes different aspects of formal systems and considers the possibility of implementing them in *digital computers*. In this context, *digital* is a description for any system that is *self-contained*, *perfectly definite* and *finitely checkable*. "All formal systems are digital in this sense" (Haugeland, 1981, p. 4). A *computer* on the other hand can be understood as a "physical device [...] which automatically manipulates the tokens of some formal system according to the rules of that system" (p. 5). Combined, this means that "any standard digital computer can, with appropriate programming, formally imitate any automatic formal system yet discovered" (p. 6). With these definitions in mind, Haugeland concluded that "the basic idea of cognitive science is that *intelligent beings are semantic engines* - in other words, automatic formal systems with interpretations under which they consistently make sense" (p. 15). Through *formal imitation*, a digital computer could simulate *any* semantic engine - if and only if it has enough resources and the right program. This formal imitation however would also guarantee *semantic imitation*, making it fundamentally possible that humans and computers "be merely different manifestations of the same underlying phenomenon" (p. 15) - something that AI researchers love to read.

Haugeland however does not end his article here, but continues to reject this thought as quickly as he developed it. He does so by presenting the two major types of approaches that aim to refute the idea of realizing human-level intelligence in digital systems: on the one hand, the *poor substitute strategy*, which argues that semantic engines will never "have the full range of common sense and values of people" (Haugeland, 1981, p. 16), that there is a fundamental impossibility in semantically imitating humans. On the other hand, there is the *hollow shell strategy*, which allows AI to go one step further and admits for imitating semantic engines in digital systems, but argues that any non-human agent may *act* like a human, yet will always lack a certain feature or property. This deficiency in for example consciousness, caring or *original intentionality* will distinguish it from really *being* human. Leaving the poor substitute strategy and manifestations of consciousness and caring to other papers, we will investigate Dennett's promising answer to the rejection of intentionality in artifacts.

3 The Intentional Stance

Intentionality can be defined as "the power of minds to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob, 2014). According to some philosophers (including Haugeland), intentionality can be *original* or *derivative*, either the product of an own mind or applied to an agent by other minds. If intentionality is solely applied, derived from the intent of an other agent, the 'receivers' thus do not necessarily need to have a mind on

their own. According to contemporary philosopher Daniel Dennett on the other hand, there is no difference between original and derivative intentionality, intentionality itself is merely a description of behavior. In his *Intentional Systems Theory*, Dennett (1971, p. 1) aims to describe the way we "interpret, explain, and predict the behavior" of other agents. To do so, it proposes three interpretation *stances* that an observer can take when describing the behavior of an agent, ordered by the amount of intent that gets ascribed to it. The most basic layer is described by the *physical stance*. By adopting it, the analysis of an agent's behavior is strictly bound to describing it through the laws of physics. For most "things that are neither alive nor artifacts, the physical stance is the only available strategy" (Dennett, 2011, p. 2). The next higher level of interpretation follows from taking the *design stance*. Artifacts' behavior can be described and predicted through their designed functionality: If something is made for a certain purpose and it works as designed, its behavior can be derived from this design. The last, most abstract level of describing behavior can be obtained by taking the *intentional stance* toward an agent. When taking this stance, "the designed thing is treated as an agent of sorts, with beliefs and desires and enough rationality to do what it ought to do given those beliefs and desires" (p. 3). This means that "anything that is usefully and voluminously predictable from the intentional stance is, by definition, an intentional system" (p. 1).
Dennett argues that by treating each other as intentional systems, using attributes such as beliefs and desires to govern interaction and generate anticipations, we "are similarly finessing our ignorance of the details of the processes going on in each others skulls" (p. 5). Through this shift of attention from actual inner workings of other people (or what we may call 'the mind') towards what we can anticipate to be their behavior according to what *ought to be* their believes and desires, humans can effortlessly adapt to new situations, not needing to fall back to any representations of schemes or scripts as proposed by researchers and philosophers of GOFAI.

4 A Continuum of Intentionality

Describing the behavior of artifacts in terms of goals or desires rather than their functionality or even physical properties is of fundamental importance to understanding them. Many researchers thus accept the idea of taking the intentional stance towards an agent - that is if two limitations are made: 1) attributions of this kind are made of derived intentionality and 2) thus are to be understood metaphorical and not literal. Dennett (2011, p. 8) however argues that 1) "there is no principled (theoretically motivated) way to distinguish 'original' intentionality from 'derived' intentionality"and that 2) "there is a continuum of cases of legitimate attributions, with no theoretically motivated threshold distinguishing the *literal* from the *metaphorical*."

Some simple artifacts (such as painted signs) indeed can be seen to not have any meaning besides the one they get through their functional role in *our* practices. But more sophisticated artifacts such as autonomous robots who function "without any direct dependence on [...] their creators, and whose discriminations give their internal states a sort of meaning to them that may be unknown

to us and not in our service" (p. 8). This opinion is even shared by AI critic Hubert L. Dreyfus (2007, p. 20) who states that "our sensitivity to relevance depends on our responding to what is significant for us given our needs, body size, ways of moving, and so forth, not to mention our personal and cultural self-interpretation." The *Dasein* of a robot would be different from our human *Dasein*, leading to goals, desires and beliefs that do not necessarily match our own - or maybe not even match our understanding of those concepts. And, as Dennett argues, "if that is not original intentionality, it is hard to say why not" (Dennett, 2011, p. 9). To underline this idea of what we might call *necessary intentionality* in complex agents, Dennett concludes with the idea that even the 'original' human intentionality must have evolved over generations of ancestors with "simpler cognitive equipment" (Dennett, 2011, p. 9) - that is, if intentionality is not assumed to be a God-given property.
So where to draw the line between 'original' and 'derived' intentionality? Dennett (2011, p. 9) remarks that "the intentional stance works (when it does) whether or not the attributed goals are genuine or natural or really appreciated by the so-called agent" and that this tolerance is the key to understanding *genuine goal-seeking.* It provides a neutral perspective for describing the behavior of rational as well as simple agents, for agents seeking their own good and agents 'just seeking' - and we do not even have to discriminate between the two. So instead of distinguishing original and derived intentionality, Dennett introduces the notion of the *order* of an intentional system: first order intentional systems are those agents we assign simple beliefs or desires to. A plant growing towards the light does so because of its *desire* for more energy. A second order intentional system on the other hand can be assigned beliefs about beliefs or beliefs about desires and so on. So on a third order level, a plant redirecting most of its energy to growing tall *wants* to escape the dense leafage of other plants which it *believes* will also *desire* for their share of solar energy. Through this order model, we see that the

This approach of assigning different orders of intentionality to an agents also allows for breaking down complex systems into fundamental intentional units with certain basic beliefs or desires. According to Dennett, eventually these units will become so simple that they reach "a level at which the residual competence can be accounted for directly at the design stance" (p. 11). Through this idea of *homuncular functionalism,* AI research would be able to construct intentional systems by combining basic functional units that do not need to have mental capacity of any sort

5 Objections considered

No theory goes unchallenged, so in this section five of the most fundamental arguments against Dennett's Intentional Systems Theory are investigated and attempted to be refuted.

1. **When do we actually start speaking of Intelligence?**

   Philosophers and psychologists such as David Premack, Donald Davidson and Robert Brandom tried to use Dennett's intentional order model in order to determine "the necessary and sufficient conditions for true believers" (Dennett, 2011, p. 11). Their approaches contain setting up a minimum order of intentionality for ‘intelligent’ beings (at least 2nd order intentionality) or the need for the ability to make thought explicit. Dennett however proposes to turn the issue inside-out, making ‘true’ cases of genuine belief or ‘intelligent’ rather “limiting cases, extreme versions, of an underlying common pattern” (Dennett, 2011, p. 12). He argues that even in the most cases of decision-making in humans, an explicit representation of relevant beliefs and desires is not present. Often we unknowingly - or unconsciously - incorporate certain propositions in our reasoning which we later cannot report on. So if we even cannot make our own thoughts and intentions explicit at all times, how do we expect to make explicit the intentions of other agents or artifacts? And with regard to the minimum level of intentionality order, Dennett gives some situational examples where non-human animals and relatively simple artifacts - agents we commonly do not believe to be intelligent - can be assigned with at least 2nd or even 3rd order intentionality. As he has argued before: There is no - and cannot be any - theoretically motivated threshold in the continuum of intentionality orders.

2. **Arguments of Blockhead and the Martian Marionette.**

   The Blockhead (Block) and Martian Marionette (Peacocke) thought experiments are examples of arguments that introduce artificial agents who appear to be human through their behavior but turn out to be no more than ‘dumb’ machines when looking behind the curtains. The Blockhead is thought of as a clone of a human but stripped of all features that are commonly assumed to define our intelligence. So when interacting with others, he actually ‘does not know what he is doing’. Dennett however refuses to see this as a valid argument, depreciating it as a philosophical zombie, a concept which many philosophers claim to be logically incoherent and thus impossible. Dennett even goes one step further and leads the idea of such zombies ad absurdum by introducing zimboes, 2nd order intentional philosophical zombies, who “think they are conscious, think they have qualia, think they suffer pains they are just ’wrong’, in ways that neither they nor we could ever discover!” (Dennett, 1996, p. 322). Now either everyone could be a zimboe, or, if we assume we are not, nobody can.

   The Martian Marionette on the other hand is accepted as a valid argument, but Dennett does not see it as a counter example to his theory. The Martian Marionette is, like Blockhead, only a human-like shell but is controlled by a super computer on planet Mars. Dennett however argues that the intentional system theory does not specify - nor need to specify - *where* the reasoning of an agent is performed. In this case, the intentionality of the Marionette thus should not be assigned to the human-like body here on Earth but rather to the computational system on Mars. And if that system controls more than one (pseudo-)agents, not the actual computer but rather the *program* running on it should be assigned intentionality. As Dennett puts it, the Marionette “simply keeps his (silicon) brain in a non-traditional location” (Dennett, 2011, p. 15).

3. **The Giant Conversation Look-up Table Objection.**

The Giant Look-up Table Objection postulates an ‘intelligent’ computer system that succeeds in the Turing Test, providing answers to the test-subject’s question that cannot be distinguished from human answers. When however the curtain is lifted, it turns out to simply be an ingeniously programmed giant look-up table that selected the best reply to each answer from a fixed set of possible answers. And simply following programmed rules to select item from a given database does not appear intelligent to us. Besides hinting at the practical impossibility of such a system, Dennett answers this objection by an analogy to the Martian Marionette argument: The actual ‘intelligent’ part in this example is not the look-up table itself but the programming that generated it. Nature creates sometimes ingenious solutions to evolutionary problems, but they are result of a sometimes long process of trial and error. Dennett thinks it to be impossible that a system like the giant look-up table developed by evolution through innumerable blind trial and error selections. He proposes such a system to be necessarily hand-crafted, being a product of intelligent design. Then he asks, “Why should it matter when the cogitation is done, if it is all designed to meet the needs of a time-pressured world in an efficient way?” (Dennett, 2011, p. 18) and argues, similar to the Marionette case, that the system simply does its thinking in ‘a strange time’.

4. **Who says that Humans are rational?**

   Until now we only discussed agents that appear to exhibit human-like intelligence but actually are proposed to be not rational at all. But there are also arguments from the exact opposite side of the spectrum: How can we prove that humans are actually rational agents? By applying intentionality to other humans and ourselves, we might just treat people much more rational than they really are. Dennett replies to this objection with two - what he claims to be - facts: That “human behavior is simply not interpretable except as being in the (rational) service of some beliefs and desires or other” (Dennett, 2011, p. 19) and that there cannot be *any* stable interpretation of mind if all behaviour is irrational.

   Aside from whether one actually acknowledges the validity of these ‘facts’, there are two other conceptual problems that arise from this issue: How do we actually know that our intentionality assignment is reliable? Dennett would argue that intentionality is that what we assign, but why not assign a rock the desire to hit the ground, using gravitational forces to realize his intend? This leaves us with two principal options: Either everything can be assigned intentionality - rendering the idea of intentionality a useless approach to the philosophy of mind; or we again need some measurement or discrimination of actual intentionality that is independent of our assignment as proposed by Dennett. With regard to this point, Dennett might not be as resolute as he wishes to be, leaving his theory vulnerable at a critical point of fundamental argumentation.

5. **Issues with the last Homunculus/Compunculus.**

   As a last point of criticism, we use exactly this vulnerable spot and investigate an objection to the soundness of his order model of intentionality: Dennett claims that with the different orders of intentionality which can be applied to an agent, its behavior can be explained by a number of simpler sub-systems - until they are so simple that they can be described by taking the design stance rather than the intentional stance. Dennett then claims that these systems can be reproduced by machines and postulates a bottom-up approach to creating Artificial Intelligence. Many philosophers claim that here Dennett fell victim to a logic fallacy called *first step fallacy*: Even if assuming that the complexity of the sub-systems decreases exponentially, there will always be an potentially unbridgeable distance between the smallest level of ascribed intentionality and no intentionality, the point where an agent becomes merely an intelligent design (Jacob, 2014). This issue is inherent to the approach of *Homunculus Decomposition* (*Compunculus Decomposition* for computers) to explaining intelligence through higher-level intelligence governing systems.

Dennett does not reply to this objection, so let us try to give a possible reaction here: Dennett does not exclude the possibility to describe the behavior of artifacts through the basic physical stance - it just is much more complex and often irrelevant to make them predictable. The same is true for intentional systems: They still can be described using the design stance or even physical stance - even if that gets increasingly elaborate. This means we do not necessarily need a clear line between 'extremely simple but intentional' and 'purely functional'. Already the simple intentional ones can be attempted to be described using the design stance. So even if that distance cannot be crossed theoretically, in practise we just proceed to build the most simple intentional subsystems.

6 Conclusion

Dennett complained during a TED talk in 2003 that most people think of themselves as experts of consciousness, that many people think that being conscious tells them all there is to know about consciousness itself (Dennett, 2003). So when proposing a rather counter-intuitive theory such as his Intentional System Theory which claims that even our own intentionality is a result of being assigned with it by others - and that through analogy many other things we think of being non-intelligent may also have intentionality - many first reactions will be refusal of such a theory. But when submerging deeper into his model of argumentation and its potential consequences for the hopes of future AI research, Dennett's theory becomes more and more relevant. By refuting the objection that assigned intentionality necessarily is derived from the observer does Intentional Systems Theory allow more sophisticated artifacts such as computers and robots to 'have' intentionality, in the more traditional terms; to reject the claim that robots can never have 'X' where X is intentionality, in the terms of Haugeland.

Intentionality however is only one of the features we want an Artificial Intelligent System to have - and maybe it not even is an exclusive feature of the mental at all. Concepts as consciousness, self-awareness, intrinsic motivation or care are considered to be equally important. Dennett's model thus only re-opens the theoretical possibility of one of those domains, leaving the others untouched. In this regard his theory should probably more be seen as a challenging new approach to understanding the working of the human mind, setting a new course for future research that should - and most likely will - be followed closely by the workers of AI.

Words: 3458

References

Dennett, D. (1971). Intentional Systems. *Journal of Philosophy*, *68*(February), 87–106.

Dennett, D. (1996). *Darwin's dangerous idea: Evolution and the meanings of life.* New York: Simon & Schuster.

Dennett, D. (2003). TED Talks - Dan Dennett: The Illusion of Consciousness. *TED Talks. [Video File].* Retrieved from `https://www.ted.com/talks/dan_dennett_on_our_consciousness`

Dennett, D. (2011). Intentional Systems Theory. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind.* Oup Oxford.

Dreyfus, H. L. (2007). Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian. *Philosophical Psychology*, *20*(2), 247–268.

Haugeland, J. (1981). Semantic Engines: An Introduction to Mind Design. In J. Haugel (Ed.), *Mind design.* MIT Press.

Jacob, P. (2014). Intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014 ed.). Retrieved from `http://plato.stanford.edu/archives/win2014/entries/intentionality/`