# Does understanding require a body?

Wouter Bouvy

July 13, 2009

## Contents

# 1 Introduction

> "It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. That process could follow the normal teaching of a child. Things would be pointed out and named, etc."
> (Turing, 1950[3])

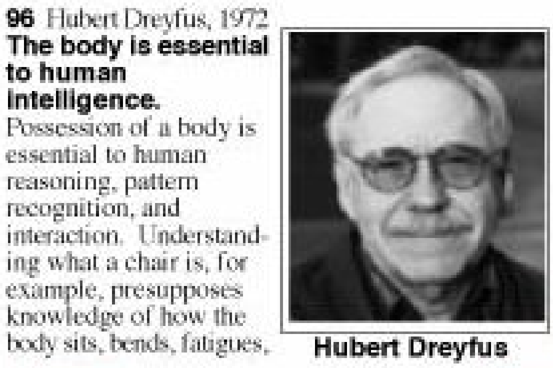
While we exist, we are continually solving problems. Most of these problems can be generally defined as a variation on the question 'what do I do now?'. This definition is broad enough to allow for both the complex problems and the simple things we might not even consider problems, like the response to an acoustic stream such as "could you pass the salt?".

In 1972, Dreyfus wrote a book about these issues, as a critique against the theory of artificial intelligence we nowadays refer to as 'Good old fashioned Artificial Intelligence'. In his book, he explained why we need a grasp of the context of problems in order to understand their meaning and determine the appropriate response.

To give a meaning to these problems is something that, according to Dreyfus, cannot possibly be done by a symbol system. There is no general method, consisting of clear and simple steps, which a person, creature or system can apply to the parts of the problem to discover the right course of action; its solution.

In this article I will see if Dreyfus' theory will fit within the 'Brain in a vat' theory[1] and look at a counter argument by Pylyshyn, and see what could be derived from these theories about the question whether machines need context.

# 2 The body is essential to human understanding




Hubert Dreyfus has argued in his book[1] that a body, or more specifically, a means of interacting with the environment, is essential for human intelligence. In this argument, considerable importance is given to the matter of context. Context is used as a term covering a large, wide-ranging phenomenon, which includes things like situation, background, circumstances, occasion, etc. The human mind always operates', if that is the right word, within a setting which permeates whatever it is it operates on.

The three basic terms, in Dreyfus' vocabulary, are 'context', 'situation', and 'background'. Dreyfus employs 'context' and 'situation' as nouns, and their importance is highlighted by the fact that they are entries in the index of the book. 'Background' does not usually stand alone, being often used in the phrase "against the background of", or appearing in expressions such as "the background

[1]Comparable to Descartes' argument in 'Meditations on First Philosophy' that he cannot trust his perceptions on the grounds that an evil demon might be controlling his every experience.

of problem solving, language use, and other intelligent behavior" and does not appear in the index of the book.

## 2.1 Context

Context seems to be defined as the settings in which humans encounter facts, problems, etc. The most important aspect of context is that one cannot define it by rules, nor represent it in any way. It does not consist of a finite set of features, or a finite set of information.

As we are never confronted with anything outside some context, context is a basic element of our mental lives. It is impossible to give a definition of a word without the context it is in, as many words have several definitions, let alone translate the single word into another language. Every possible use of the word is likely to have some variation in the target language.

Also, every problem or task has a context[2] that needs to be taken into account. Lingual interaction cannot work without some context for interpretation (e.g. slang, rhetorical questions, figurative meanings). Any fixed meaning (using rules, knowledge representation, etc) that may be attached to a word, task or sign has to be mixed with some ingredients coming from the context in some way.

The process of understanding consists of determining the contribution of context. In other words, one must discover the relevant aspects of the context and find the appropriate representations of those aspects. These representations would need to be combined with the meaning assigned to the word to infer the understanding.

## 2.2 Situation

Situation seems to be defined as a collection of states of the world. A situation cannot, in principle, be turned into representations that we can use to determine the proper response to a problem. Dreyfus states that situations are holistic; the only moment when the individual parts of the situation have meaning is when they are in the situation. The parts do not have an intrinsic value, nor do they exist outside the situation. By representing the situation as a symbol system, you evaluate the situation per part, which is not possible.

Also, he states, one needs to take into account that a situation already contains expectation, goals, etc of a system. To compute the contribution the situation has to the environment is therefore impossible. The contribution of the environment is already influenced by the expectations and goals of the situation.

## 2.3 Background

Background is a vague concept in Dreyfus' work; it is never clearly defined, and probably cannot be clearly defined. He attempts to define it by stating that background is "something like an ultimate context [...] the even broader situation call it the human life-world". He attempts to argue that this is something that cannot be defined, comparable to the argument that the collection of all sets is too large to itself be a set.

Dreyfus's major references are Heidegger and Merleau-Ponty. In the Heideggerian framework, the background is provided by a particular state of being in which an 'involved' human being (not a simple creature) deals with a situation that is already filled with meaning. The background will contain the necessary equipment for dealing with this situation.

He relies on the idea of a bodily origin of meaning as stated by Merleau-Ponty. This basically states that action comes before reflection, body before mind. "He [Merleau-Ponty] argues that it is the body which confers the meaning", writes Dreyfus. The logical world of context-free rules guide

[2] I thought, at first, that some mathematical problem might be considered to not require context, as they are universally applicable, but then, mathematics itself is most likely a context as well.

the beginner, but for more experienced entities these rules are transformed into a set of generalized bodily habits or skills by a learning process seated in the body.

Dreyfus does not want to reduce background to the neurofysiological description. Dreyfus does not want to accept the possibility of any intentional content being disconnected from involved coping. Although he does allow for the existence of derivative intentional content, which is dependent on theoretic understanding acquired earlier in life. This could not, of course, exist without language, so there is a limited domain of entities which will have this type of intentional content.

## 3 Brain in a vat theory

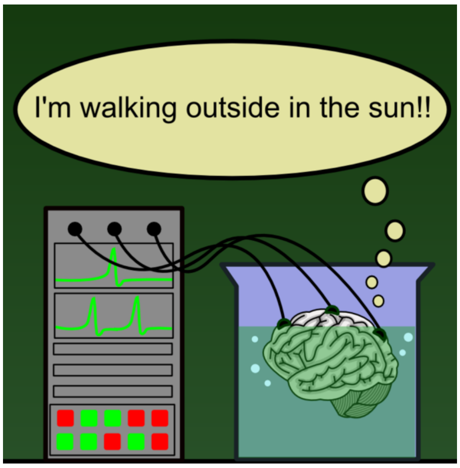



Image a science fiction story, where a mad scientist removes a person's brain from the body, suspends it in a vat of life-sustaining liquid, and connects its neurons by wires to a supercomputer. The supercomputer would provide it with electrical impulses identical to those the brain normally receives, simulating a virtual reality (including appropriate responses to the brain's own output) and the person with the "disembodied" brain would continue to have perfectly normal conscious experiences without these being related to objects or events in the real world.

This thought experiment can be used to for many theories about our interaction with the world, on the bases of the following. The brain in the vat cannot differentiate between being in the vat (and thus, the simulation) and being in the body of a person (and thus, in the real world), because all inputs, outputs and phenomenon outside the body are simulated. Because of this, you, as a person, cannot know whether you are a brain in a vat. You may experience walking, sitting, eating, reading, etc, but you will never know if the things you walk on/to/around, sit on, eat or read are actually physical objects or simulated object.

But what if the brain in the vat was made aware of the fact that it is a brain in a vat? It could never grasp the idea of being a brain in vat, unless the simulation it is in is an exact copy of the outside ('real') world. In other words, if a brain in a vat states that it is a brain in vat, it would always be stating a falsehood. What the brain is really stating is "I am what I have been convinced is a 'brain,' and I reside in an image that I have been convinced is a 'vat'."

The brain in a vat could never think about real brains or real vats, but rather about images sent into it that resemble real brains or real vats. There are two ways for the brain to state the truth: if the brain lives in the real world or if it has lived in the real world and has experienced these concepts while in the real world. In the first case, it is obviously not a brain in a vat. In the second case, if the simulation is good enough, it will never know when it was and was not in the real world, and thus cannot determine whether the concepts are from the real world. Or, if the simulation is weaker than the real world, the brain[3] would have either discovered that something

[3] Or person

was wrong with the outside world, or attributed the change to a deprecation of his own senses (especially if the simulated entities all claim nothing has changed).

This leads to the following conclusion; if you've never experienced the real world, then you can't have thoughts about it.

Dreyfus' theory might be merged with the brain in a vat theory. After all, we might be living in a simulation, be it singular [4] or shared [5], and still have all the elements Dreyfus considers necessities for intelligence and understanding.

As noted above, we would not be able to understand the external world in any way, not having any experience with it, but this is not relevant for the individual inside the simulation.

Imagine a human being, who has been living, from birth, inside a simplified simulation of reality where everything is two-dimensional. This person would not understand our three-dimensional world, but would function as a 'normal' entity inside the simulation. His understanding would only be limited by the two-dimensionality of his world.

Given a burst of inspiration, he might even be able to conceive that there could be more than two dimensions, as we can conceive more than 3 dimensions. He might even be able to reason about these dimensions in some abstract way, by attempting some reduction towards his environment, e.g. a hypersquare describing a cube, comparable to our definition of a hypercube describing a 4-dimensional object with axes of equal length, etc.

Ultimately, this person would never fully understand the world we live in, until we release him from the simulation and allow him to adapt to our three dimensions[6]. Even then, having grown up inside the two-dimensional world, the person would hold on to the things learned when growing up, and try to adapt these to the new three-dimensional world.

# 4 The body is only essential for the development of understanding

A response to Dreyfus' argument, as given by Pylyshyn in 1974[2], is that only during the development of the mind the body is critical. Basically, the body provides the necessary interaction to discover information about objects in the world and properties of the world, but once this information has been learned the interaction with the object is no longer necessary to understand the object.

This is not really a counter-argument, as it still states that the context s required, and that one needs to understand the world, but it limits the extent of the need for interaction with the environment. It can arguably be considered a specification of, rather than counter argument to, Dreyfus' theory.

A blind person will never be able to understand colors. One might explain colors in terms of emotions, or sensations like warm and cold. But this will never give the blind person a full understanding of the color, just an approximation or a definition in terms of fully-understood phenomenon. If, on the other hand, a person loses his ability to see during life, he will still be able to remember what each colors look like. This person will still be able to reason about the colors, have a favorite color and will still fully understand the color.

[4] e.g. where everything you encounter is simulated, every object or person

[5] e.g. the Matrix, where the environment is simulated, the objects are simulated, but (most of) the other people in the world are not simulated

[6] Assuming, of course, that he does not go completely mad from the transition.

This comparison will hold true for any of the five senses, losing one does not equal never having had one. This effect becomes more obvious if several senses are impaired. While someone who has lost the ability to see and hear may still communicate in roundabout ways (using touch and gestures on the skin), his understanding of the world will not be diminished. On the other hand, a person who has never had sight or hearing, will have a lot of trouble trying to understand the world. The person may learn language through the combination of holding an object and feeling someone make the gesture for the name of the object on the skin of the person, but it will always be an impaired understanding of the world.

### 4.1 Our missing senses

One might argue that our understanding of the world is far from complete. As stated above, a blind person will never understand colors. This can be extended to the entire human race. Our visual spectrum is but a small part of the entire electromagnetic spectrum. We will never be able to understand phenomenon like x-rays and microwave other than in abstractions to the limited world of our senses. And that is only the limitation of one of our senses, not the total absence of one.

There might be many phenomenons we might not be able to conceive, simply because we have no sense that approaches the definition of the phenomenon. We might know of some of these senses because we have seen them in animals, like echolocation in bats, but our implementation of sonar is a visual thing (e.g. submarine sonar, radar, etc). Others, like electroreception[7] in fish, magnetoreception[8] in birds and pressure detection in fish, are a lot more distant from our sense, and we can only approach these using meters and numbers (or visual representations).

Thus, we can question whether we should demand that our artificial intelligence understands things as we do. We are obviously not completely equipped to understand everything fully, why should we not accept the same limitation on our machines? The answer here would be that we cannot communicate sensibly with machines that perceive a completely different world than we do. If our machine misses a sense that we use regularly, we cannot describe our environment using that sense and this will severely limit communication. Thus, to ensure a somewhat fluid interaction, we will need a machine with at least some of our senses to share at least some understanding of the world.

## 5 Learning machine

Dreyfus may be correct that there may be a problem for a certain type of system (e.g., a rule-based system), but there may be other types of systems that avoid this problem. For example, Rodney Brooks' attempts to build simple insect-level intelligence or several new connectionist networks approaches might not have the issues Dreyfus assigns to rule-based systems.

Assume that we have a machine that can learn to cope with simple environments. This machine may start out as a clean slate, and learn to cope in the (simulated) environment it is growing up in. After a certain amount of time, this machine may have learned enough to be a fully functional organism inside its world, whatever that may entail in the simulation. One might even say it can 'understand' some of the simple concepts of its world.

Of course, like humans, its development is never finished. It never reaches the point where it is done[9]. But assuming there is some point at which we can consider it adequate, we can conceivably copy it entire 'mind state', so to speak, into a new machine.

From that point on they would grow to be different entities, but the second machine has arguably stepped over the requirement of environment for development. Alternatively, one might state that the machine did need the environment to develop its understanding of the world, but that this was shared with the other machine.

[7]detection of electric fields
[8]detection of magnetic fields
[9]Though, this might be possible given the simple nature of the environment.

## 5.1 Generate it?

A question that might rise from this approach, though, is "If we can copy it, can we not just generate it?". The problem with learning machines (let's assume a neural network for the moment) is that we cannot see what it has learned. All we can see are weights and connections. We can however, tweak these and see what happens. Extended to the extreme, we could generate every possible configuration (or a portion of the range of possible configurations) and see which perform adequate.

This may seem like a way to generate the low-level intelligent behavior of the machine as described above, without requiring the environment for learning. Thus, we might not need the environment for development after all. Of course, this is an over-simplification. Even a very simple entity in a very simple environment will need many neural nodes to and connections to learn to interact with the environment. For example, a 25 input, 25 unit hidden layer and 2 unit output layer neural network has 702 free parameters!

## 5.2 Inheritance of behavior

A comparable effect can be observed in the animal world, where some behavior seems hard-coded into the animals, such as hibernation, migration, nest building, mating, etc. Many more will be learned in the short time an animal grows up near its parent. Of course, given the changing nature of the environment, it does not pay for animals to inherit every behavioral trait from their parents, thus not everything is instinctual. If our environment for the machine as described above is changing, we might not want every machine to have the same learning process (by copying). If however, our environment is somewhat static (e.g. repeated tasks that do require a measure of adaptability, but only in a limited way[10]), cloning of our capable machine might be a viable option.

But here we see the problem. Our simulation might be static at some point, but eventually, we will want the simulation to approach the complexity of our (perceived as real) environment. The smaller the step from simulation to reality is, the higher the chance that the 'brain' will be able to cope with reality when it is removed from its 'vat'.

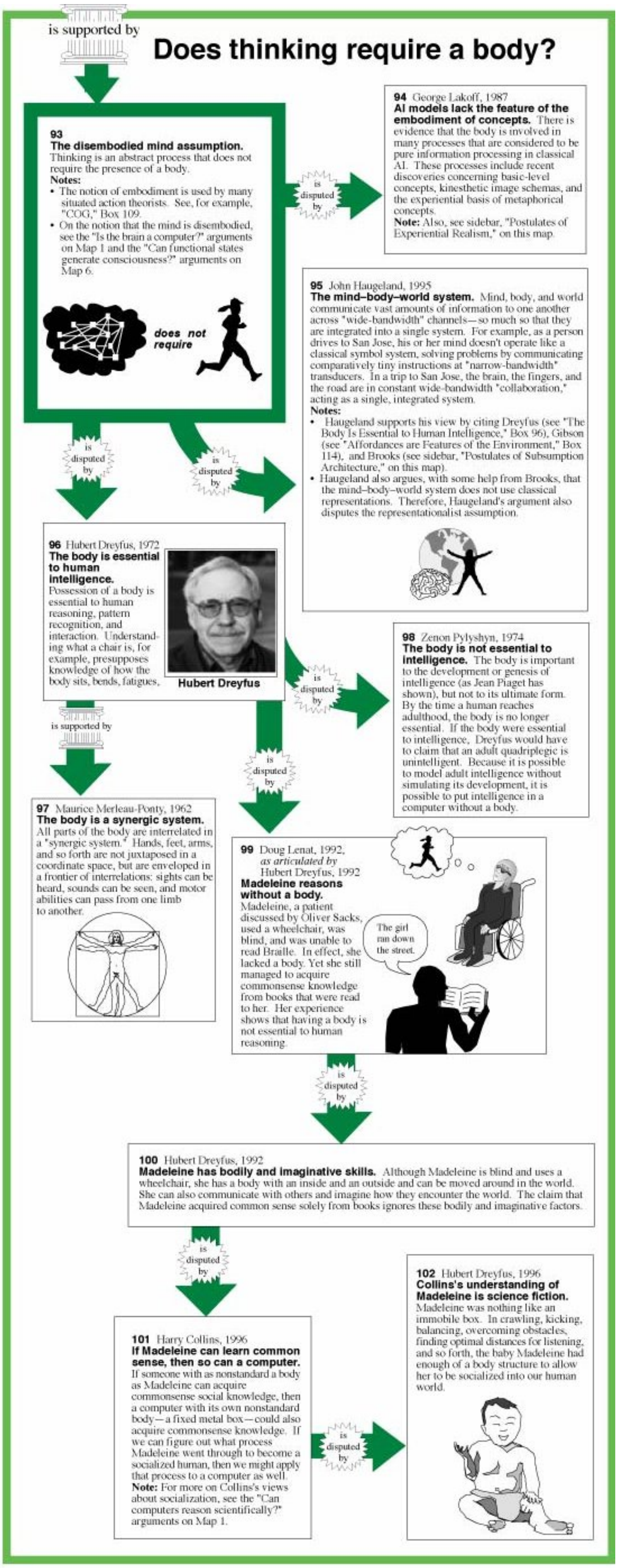
We would have to clone a limited part of the machine's understanding of the world, requiring the new machine, like animals in our world, to still learn how to cope with it themselves, in order to ensure adaptability. This would still require us to conclude that that the environment is necessary to develop an understanding of the world.

# 6 Conclusion

In this paper we have discussed Dreyfus' argument against 'Good old fashioned AI', which may also apply to newer approaches in AI. We have looked at the 'Brain in a vat' theory, and seen that Dreyfus' theory is versatile enough to allow the fact that we may live in a simulation.
We have looked at a counter argument by Pylyshyn, which states that we only need the environment to develop our understanding of the world; afterwards we have enough to deal with most problems that arise, even if we lose our ways of interacting with the environment.
Considering that Dreyfus' theory does not exclude simulation, and given the argument by Pylyshyn, we have given the idea of a machine, in a simplified simulation, that can be copied once it has developed its understanding of the world. We have seen that this can be considered a way to circumvent the requirement to learn to deal with the environment by way of a sort of inherited, 'instinctive' understanding for the new machine. The parallel with animal inheritance is obvious, although animal inheritance allows for individual learning to ensure adaptability to a changing world.

[10] Though this might not need an 'intelligent' system anyway.

# 7 Philosophical map of this theory



# References

[1] Hubert L. Dreyfus. *What Computers Can't Do: The Limits of Artificial Intelligence*. Harpercollins, 1972.

[2] Z.W Pylyshyn. Minds, machines, and phenomenology: Some reflections on dreyfus's "what computers can't do". *Cognition*, 3(5):55–77, 1974.

[3] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.