Are We Living in The Matrix?
Response to the Argument of Nick Bostrom
and the Strong AI Hypothesis

Last Update: 10th July 2008

Nick Bostrom is in the department of philosophy at Oxford. His argument can be found at http://www.simulation-argument.com/

The briefest summary of his argument is as follows: If humans survive long enough to develop the capability to create computer simulations of their history, and actually do so to a 'significant' degree, then we are almost certainly living in such a simulation.

So you can believe just one of two things: that humans will never develop such a capability, or that we are living in a simulation. In as far as the idea that humans could develop such computational power seems reasonably likely, then it becomes equally likely that we are living in a simulation. The creator of this simulation is therefore Our Creator. This has been called the first interesting argument for the existence of a Creator in 2000 years.

Actually, the computer simulations which Nick Bostrom envisages are different from those in the film "The Matrix" in a crucial way. In "The Matrix" the humans are real. It is only their experiences which have been high-jacked by the Aliens. Their human bodies, and their capacity for subjective experience and consciousness, are all real. "The Matrix" therefore presents no philosophical problem as regards its implementation, only a technological one.

In contrast, Bostrom's simulations are simulations of the entire human. This presents a very major philosophical problem. The strong AI hypothesis[1] is assumed. Specifically, what Bostrom refers to as an "attenuated version of substrate independence" is assumed. This means that, "it would suffice for the generation of subjective experiences that the computational processes of a human brain are structurally replicated in suitably fine-grained detail, such as on the level of individual synapses". In other words, any computing device which replicates the computational steps of a human brain would automatically give rise to the same subjective experiences and consciousness. The simulation, or rather the component of the simulation representing one organism, would have self-awareness and be convinced that it was alive. Even to the philosophically uneducated (such as myself) this is quite an assumption.

The counter-argument, that the strong AI hypothesis is false, has been made by John Searle in "Minds, Brains, and Programs," in *The Behavioral and Brain Sciences*, vol. 3, 1980, Cambridge University Press. This paper can be found at, http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-00/bbs.searle2.html

[1] I use the term "strong AI" in the sense of John Searle, meaning a computing device that can actually think and has a mind, with all the associated self awareness, subjectivity, etc., associated with minds. The "strong AI hypothesis" is that a computer which implemented computational steps isomorphic with those of the synapses of a living brain would thereby achieve strong AI.

This paper includes the famous “Chinese Room” argument, which counters the claim of substrate independence. Understanding, subjective experience, self-awareness, consciousness and the rest are not generated automatically by a dumb program implemented on an arbitrary computing platform (argues Searle).

But the thing that stunned me when I read this paper was Searle’s observation that those who believe in strong AI as a way of constructing minds are actually dualists. A dualist[2], such as Descartes, believes that mind and body are quite distinct and that humans consist essentially of both. At first glance, therefore, it seems most odd to accuse proponents of strong AI of being dualists. They believe you can create minds out of dumb matter, arranged in the form of computers. Surely that is exactly the opposite of a dualist?

The issue is expressed most clearly in terms of the strong AI fraternity’s faith in substrate independence. They believe that the essence of mind is captured by a program. The two components of their dualistic world are hardware and software, for which read “body” and “mind”. Here are Searle’s own words,

“…this residual operationalism is joined to a residual form of dualism; indeed strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter. In strong AI (and in functionalism, as well) what matters are programs, and programs are independent of their realization in machines; indeed, as far as AI is concerned, the same program could be realized by an electronic machine, a Cartesian mental substance, or a Hegelian world spirit. The single most surprising discovery that I have made in discussing these issues is that many AI workers are quite shocked by my idea that actual human mental phenomena might be dependent on actual physical-chemical properties of actual human brains. But if you think about it a minute you can see that I should not have been surprised; for unless you accept some form of dualism, the strong AI project hasn't got a chance. The project is to reproduce and explain the mental by designing programs, but unless the mind is not only conceptually but empirically independent of the brain you couldn't carry out the project, for the program is completely independent of any realization.”

On reflection, the dualism of mathematicians is apparent and familiar. It has been said that, irrespective of any avowed philosophy, the working philosophy of every mathematician is Platonism. It is hard to see how anyone could spend time and effort in its study if they did not believe that its subject matter existed. And since the objects of study in mathematics clearly do not exist in the world of spacetime and matter, it follows that all mathematicians implicitly believe in an alternative Platonic realm. Hence they are dualists in this sense. The idea of a program divorced from its hardware also exists in this same Platonic world. So, Searle’s accusation that the strong AI fraternity are dualists appears to be correct. On the other hand, and by the same token, it appears hard to avoid such dualism. So has Searle merely scored an own-goal?

However, whether or not strong AI is implicitly dualist seems independent of whether the strong AI hypothesis is correct. If it were correct, then the self awareness and so on

[2] As opposed to a duellist, of course, who has other means of argument.

would appear to reside in the Platonic realm also. But what is the ‘self’ of which this program becomes aware? And how could external observers, such as ourselves, become cognisant of its consciousness? The required communication channel can only be via the world of sticks and stones. Also the definition of ‘self’, and the evolution of self awareness, would seem most directly accomplished via the world of ponderable matter. So it may be that dualism is indeed necessary, and that the phenomena of self awareness, consciousness, etc, are best understood in terms of the relationship between mind (program) and body (hardware). But both parts of the dual world are essential. This is Searle’s position, I think.

So, how does all this relate to Bostrom’s argument that we might be living in a simulation? Simply that it challenges the notion that “living in a simulation” makes any sense at all. That is, it (perhaps) makes no sense so long as we regard the simulation, as Bostrom does, as a purely digital affair – a formal program independent of substrate.

On the other hand we could envisage, if you will, an ‘analogue simulation’. For this our creator would need to create the physical material out of which we are made, in addition to the required ‘software’. But this becomes synonymous with the creation of our universe, no different or in any way less real than it actually is. This is no longer a simulation at all. It is reality. So the realisation that both parts of the dualism, mind and body, are required saves us from being too readily simulated.

This exposes the weakness of Bostrom’s argument. By assuming that a software simulation is sufficient, the creation of sentient beings is made to seem too easy. It becomes very reasonable to suppose this would be achieved by post-humans. But if, in truth, specific hardware implementations are crucial, then the ‘simulation’ requires creation of a new, artificial, physical universe containing the simulated life. It is no longer so reasonable to suppose, arbitrarily, that such a thing will become possible. And hence there is no particular reason to regard this universe as being simulated.