**Consciousness, Agents and the Knowledge Game**

Luciano Floridi
Faculty of Philosophy and IEG, Computing Laboratory, Oxford University.
Address for correspondence: Wolfson College, OX2 6UD, Oxford, UK; luciano.floridi@philosophy.oxford.ac.uk

**Abstract**

This paper has three goals. The first is to introduce the "knowledge game", a new, simple and yet powerful tool for analysing some intriguing philosophical questions. The second is to apply the knowledge game as an informative test to discriminate between conscious (human) and conscious-less agents (zombies and robots), depending on which version of the game they can win. And the third is to use a version of the knowledge game to provide an answer to Dretske's question "how do you know you are not a zombie?".

**Keywords**

Artificial agents, consciousness, inferentialism, knowledge game, "muddy children" theorem, "the three wise men" theorem, zombies.

"Silently Peyton weighed his opponent. It was clearly a robot of the very highest order. […]
'Who are you?' exclaimed Peyton at last, addressing not the robot, but the controller behind it. […]
'I am the Engineer.'
'Then come out and let me see you.'
'You are seeing me'. […]
There was no human being controlling this machine. It was as automatic as the other robots of the city – but unlike them, and all other robots the world had ever known, it had a will and a consciousness of its own."
A. C. Clarke, *The Lion of Comarre*, 1949.

## 1. Introduction: how do you know you are not a zombie?

Consciousness is one of those fish we seem to be unable to catch, much like intelligence. We recognise its presence, traces and effects, but its precise nature, workings and "location" still escape our grasp. Tired of ending up empty-handed, some philosophers have recently tried to approach the problem of consciousness indirectly. If you can't hook it, try to corner it. To this new approach belongs a series of mental experiments involving the possibility of conscious-less agents (see for example Symposium [1995]).

Imagine three populations of agents: robots (conscious-less artificial agents), zombies (conscious-less biological agents) and humans (conscious biological agents). I shall say more about the first two types of agents presently. There is of course a possible, fourth population, that of conscious artificial agents like *The Engineer* in Clarke's story, but I shall not consider it in this paper. At the moment, the assumption is that you, a human, are neither a robot nor a zombie and that you know it. This much is granted. The question is *how you know it*.[1]

Dretske [2003] phrases the problem neatly: "I'm not asking whether you know you are not a zombie [or a robot, my addition]. Of course you do. *I'm asking how you know it*. The answer to that question is not so obvious. Indeed, *it is hard to see how you can know it*. Wittgenstein (1921/1961: 57) didn't think he saw anything that allowed him to infer he saw it. The problem is more serious. There is nothing you are aware of, *external or internal*, that tells you that, unlike a zombie, you are aware of it. Or, indeed, *aware of anything at all*. (my emphasis)".

[1] Compare this with the skeptical problem about propositional justification: one may be justified in believing that *p*, without this warranting that one is also able to know that one is justified (Alston [1986]). As Descartes saw, one may try to get out of this predicament by making sure that the test (for him, the method of doubt) run to check whether one is justified in believing that *p* brings out one's knowledge that one is justified in believing that *p* (Floridi [1996]).

Whatever your answer to Dretske's question is, it will cast some light on your conception of consciousness, but before we embark on any further discussion, let me introduce our *dramatis personae*.

Artificial agents are not science fiction but advanced transition systems capable of *interactive*, *autonomous* and *adaptable* behaviour.[2] *Interactivity* means that artificial agents and their environments can act upon each other effectively. *Autonomy* means that the agents can perform internal transitions to change their states teleologically, without direct responses to interactions. This property imbues agents with a certain degree of complexity and decoupled-ness from their environments. *Adaptability* means that the agents' interactions can change the transition rules by which they change states. This property ensures that agents might be viewed, at a given level of abstraction (Floridi and Sanders [2004]), as learning their own mode of operation in a way that depends critically on their past interactions and future goals.

According to Dretske – and indeed, rightly, to anyone using this thought experiment – zombies are agents that lack consciousness of *any kind*. Now, this initial definition needs to be refined, since there are four main senses in which an agent Ag can be said to possess or lack consciousness, not all of which might be in question here.

Ag may be *environmentally conscious* if

e.1) Ag is not "switched-off", e.g., if Ag is not asleep, comatose, fainted, anaesthetised, drugged, hypnotised, in a state of trance, stupor, catalepsy, or somnambulism and so forth;

or (depending on one's approach this may be a disjunctive or)

e.2) Ag is able to process information about, and hence to interact with, Ag's surroundings, its features and stimuli effectively, under normal circumstances.

Animals, including human agents are normally said to be conscious in the (e.1) or (e.2) sense. But Ag may also be *phenomenically conscious* if

*p*) Ag experiences the qualitative, subjective, personal or phenomenological properties of a state in which Ag is. This is the sense in which Nagel [1974] famously speaks of being conscious of a certain state as having the experience of "what it is like to be" in that state;

[2] For an introduction to agents and distributed systems see Wooldridge [2002].

or (and this is at least an inclusive "or", and at most a misleading place-holder for a double implication, more on this in section seven) Ag may be *self-conscious* if

*s*) Ag has a (second- or higher-order) sense of, or is (introspectively) aware of, Ag's personal identity (including Ag's knowledge that Ag thinks) and (first- or lower-order) perceptual or mental experiences (including Ag's knowledge of what Ag is thinking).

All four states are informational in character: *e-consciousness* is externally oriented and first-order, whereas *p-consciousness* and *s-consciousness* are internally oriented and (at least in the case of *s-)* second- or higher-order.

These distinctions are useful to clarify Dretske's question. For it may be unfair to interpret Dretske as saying that zombies lack consciousness even in the environmental sense. Perhaps some zombies can be conceived as being both "switched-off" (e.1) and incapable of any effective informational interactions with the environment (e.2). But then, not only might it be easier for us to say how we know that we are not *that sort* of zombies, it would also be far less interesting to investigate how we know it. So I suggest we restrict Dretske's claim to saying that zombies are biological agents almost like us, but for the fact that they lack both *p-consciousness* and *s-consciousness*. Indeed, Dretske seems to embrace the same analysis: "The properties you are aware of are properties of – what else? – the objects you are aware of. The conditions and events you are conscious of – i.e., objects having and changing their properties – are, therefore, completely objective. They would be the same if you weren't aware of them. *Everything you are aware of would be the same if you were a zombie*.[footnote] *In having perceptual experience, then, nothing distinguishes your world, the world you experience, from a zombie's*. (my emphasis)".

According to our refined definition, zombies are biologically embodied cognitive systems capable, like us and some artificial agents, of *some kind* of first-order, informational and practical interactions with their environment. The "kind" does not have to be anything human-like, it only needs to be indistinguishable (and not even intrinsically, but only by us) from an ordinary human agent's way of dealing fairly effectively with the world.

The idea is not new. There is a delightful passage in Spinoza' *On the Improvement of the Understanding*, for example, where he compares sceptics to automata (zombies, in our more refined vocabulary). The passage is worth quoting: "If there yet remains some sceptic, who doubts of our primary truth, and of all

deductions we make, taking such truth as our standard, he must either be arguing in bad faith, or we must confess that there are men in complete mental blindness, either innate or due to misconceptions – that is, to some external influence. Such persons are not conscious of themselves. If they affirm or doubt anything, they know not that they affirm or doubt: they say that they know nothing, and they say that they are ignorant of the very fact of their knowing nothing. Even this they do not affirm absolutely, they are afraid of confessing that they exist, so long as they know nothing; in fact, they ought to remain dumb, for fear of haply supposing which should smack of truth. Lastly, with such persons, one should not speak of sciences: for, in what relates to life and conduct, they are compelled by necessity to suppose that they exist, and seek their own advantage, and often affirm and deny, even with an oath. If they deny, grant, or gainsay, they know not that they deny, grant, or gainsay, so that they ought to be regarded as automata, utterly devoid of intelligence."

When Dretske asks how you know that you are not a zombie (or one of Spinoza's automata), he is not wondering how you know that you are either *p*- or *s-conscious*. For Dretske is mainly concerned with perception and "the attitudinal aspect of thought", as he writes. His question about consciousness is "how one gets from what one thinks – that there is beer in the fridge – to a fact about oneself – that one thinks there is beer in the fridge. What you see – beer in the fridge – doesn't tell you that you see it, and what you think – that there is beer in the fridge – doesn't tell you that you think it either." In other words, what we are looking for is not how we know that we are *s-conscious* – although, if we were zombies, indicating how we know that we are *s-conscious* would highlight an important difference between us and them – but how we know that we are *p-conscious* of something when we are perceiving that something. To quote Dretske once more "[…] What we are looking for, remember, is *a way of knowing* that, unlike zombies, we are conscious of things. (my emphasis)". Dretske's question *does not exclude* any reference to *s-consciousness* in principle, but it is meant to address primarily *p-consciousness*.

Now that we have defined the relevant agents and types of consciousness, a final point in need of clarification concerns what we are in principle asked to provide in terms of an answer. In the article, it is clear that Dretske (again, rightly) understands his "how" question in two slightly different ways:

Q.1) as a question that can take as an answer a specification of how one *actually* knows that one is "conscious of things" and hence not a zombie; and

Q.2) as a question that can take as an answer "a way of knowing that, unlike zombies, we are conscious of things", that is, how one can *possibly* know that one is not a zombie.

An answer to Q.1 could describe some actual further experience that one ordinarily enjoys when having a *p-conscious* experience, in order to show how one knows that one is not a zombie. An answer to Q.2 is less demanding, for it could describe just *a* (not *the*, and not *the only*) possible experience usable as a way of knowing and explaining how one knows that one is not a zombie. Q.1 invites the identification of the right sort of factual description or logical fact inferable from, or perhaps just somehow related to, the conscious experience of the world, whereas Q.2 invites the identification of some sort of informative test.

The distinction is crucial. We normally move from one to the other type of question when we realise that a test is probably the best way of explaining how one knows that one qualifies as a certain kind of agent. Yet, this erotetic shift may seem problematic. Let me illustrate the difficulty with an analogy.

Suppose you are a good cook and that you know that you are a good cook, but that you are asked to explain how you know it. Suppose that we agree that there is no actual special experience, somehow related to your cooking experience, that you and I would find entirely satisfactory to clarify the matter. As Dretske does in the article, we would be shifting to a Q.2-type of question: is there anything at all – not just an actual something, but maybe just a possible something – that could ever count not merely as evidence that you are a good cook but as a way of knowing that you are a good cook?

One standard solution would be to devise a satisfactory gourmet test and see whether you can pass it. Passing the test would not merely confirm your qualities as a good cook: your culinary capacities are not in question, so the test would not be informative in this respect. Nor would the importance of the test consist in its capacity (which it does have) to make you know that you are a good cook. Its importance would lie in the fact that it would provide you with a way of explaining how you know that you are a good cook. This because you have had the successful experience of cooking well *while* cooking well *and while* the former experience was positively assessed as qualifying you as a good cook, and you know all this.

Someone may not be convinced. Suppose you get full marks for your pasta but only barely pass the test for puddings. Since you are a good cook, we assume that you

pass the gourmet test overall. At this point, an observer may still object that the original question was of a Q.1-type, but that the test provides an answer only to a Q.2-type of question. In general, very few people qualify as good cooks because they have passed a gourmet test. If they are good cooks they are so for other reasons than passing an official exam. Moreover, a test can only certify the presence/absence of the property in question (e.g. being a good cook) at best, but it does not tell you in any detail what it takes to have it. The objection is that answering a Q.2-type of question fails to address the original concern.

The reply to this objection is twofold. First, we moved from a Q.1-type to a Q.2-type of question because we agreed that answering Q.1 may be impossible. If this is all our interlocutor wishes to see acknowledged, it had already been conceded. However, and this is a crucial point, by moving to a Q.2-type of question – as Dretske himself does – the test we have devised is not merely a successful way of discriminating between good and bad cooks. It is also informative (for the agent being tested) about what being a good cook means: it means going through the process you just went through (recall your good pasta and the bad puddings) and qualify because of it as a good cook. The test is not any test, but a test that concerns precisely the way you actually cook, only it examines it as a "work in progress" and in a context constrained by well-specified conditions, whereby the process is assessed as the right sort of process to qualify as a certain kind of agent. Good tests usually are informative, that is, they usually are more than just successful criteria of identification of *x* as *y*, because they examine the very process they are testing precisely while the process is occurring, and so they provide the tested agent with a way of (a) showing that he qualifies as a certain kind of agent, (b) knowing that he is that kind of agent, and (c) answering how he knows that he is that kind of agent, by pointing to the passed test and its (a)-(b) features.

All this means that, when Dretske seeks to show that there is nothing you are aware of that tells you that you are aware of anything, I shall argue that he may be right in suspecting that no answer to the "how" question understood as in Q.1 is available, but that he is too pessimistic in concluding that therefore there is no other possible answer to the Q.2 version either, i.e., no possible "way of knowing that, unlike zombies, we are conscious of things", for one can devise an informative test about consciousness such that passing it provides a possible way of explaining how we may know that, unlike zombies, we are conscious of things.

To summarise: according to Dretske, there is nothing you are aware of that tells you that you are aware of it, although you are indeed aware of it and you do know that you are, so it may be impossible to answer the question "how do you know you are not a zombie?". Now, one way to ascertain whether *x* qualifies as *P* is to set up a *P*-test and check whether *x* passes it. You know you are a car driver, a chess master, a good cook, or that you are not visually impaired if you satisfy some given standards or requirements, perform or behave in a certain way, win games and tournaments, pass an official examination, and so forth. This also holds true for being intelligent, at least according to Turing [1950]. I shall argue that it applies to *s-consciousness* as well and that, since *s-consciousness* implies *p-consciousness*, an *informative* test for the presence of the former is also an *informative* test for the presence of the latter. I agree with Dretske that mental and perceptual experiences may bear no hallmarks of consciousness or of any further property indicating our non-zombie (and non-artificial) nature. Consciousness (either *p*- or *s*-) does not piggyback on experience, which tells us nothing over and above itself. Blame this on the transparency of consciousness itself (it is there, attached to experience, but you cannot perceive it) or on the one-dimensionality of experience (experience is experience, only experience, and nothing but experience). However, I shall argue that this does not mean one cannot devise a reliable and informative test for the presence of consciousness as a way of showing that, and explaining how, one knows that one is a *p*- and *s-conscious* agent. The knowledge game is such test.

## 2. The knowledge game

The knowledge game is a flexible and powerful tool with which to tackle a variety of epistemic issues.[3] It closely resembles Turing's test and it exploits a classic result, variously known as the "muddy children" or the "three wise men" theorem, the drosophila of epistemic logic and distributed AI.[4]

[3] Two very different uses of the knowledge game can be found for example in Lacan [1988], discussed in Elmer [1995], and in Shimojo and Ichikawa [1989]. I hope to show the applicability of the knowledge game to the dreaming argument and the brain-in-a-vat or malicious demon hypothesis in another paper.

[4] The classic version of the theorem has been around for decades. It is related to the Conway-Paterson-Moscow theorem and the Conway paradox (see Groenendijk et al. [1984], pp. 159-182 and Conway and Guy [1996]) and was studied, among others, by Barwise and Etchemendy [1987] and Barwise [1988]. For some indications on its history see Fagin et al. [1995], p. 13. The social game *Cluedo* is based on it. Its logic is analysed in Ditmarsch [2000]. The *Logics Workbench* is a propositional

The game is played by a multi-agent system comprising a finite group of at least two interacting agents with communicational and inferential capacities. Agents are assigned specific states in such a way that acquiring a state *S* is something different from being in *S* and different again from knowing that one is in *S*. The states are chosen by the experimenter from a *commonly known* (in the technical sense of the expression introduced in epistemic logic: all players know that all players know that all players know… the) set of alternatives. The experimenter questions the agents about their states, and they win the game if they answer correctly. Agents can determine the nature of their state inferentially and only on the basis of the informational resources available. They cannot rely on any innate, a priori or otherwise *privileged access* (Alston [1971]). Most notably, they have no introspection, internal diagnosis, self-testing, meta-theoretical processes, inner perception or second order capacities or thoughts.[5] Since the game blocks the system from invoking any higher-order, mental or psychological *deus ex machina* to ascertain directly the state in which the system is, we test the presence of *p*- and *s*-*consciousness* indirectly and avoid the problem of dealing with these two types of consciousness by means of concepts that are at least equally troublesome.

Let me now sketch how the knowledge game will be used. In the following pages, we shall compare three types of agents: humans (agents who enjoy not only *e*- but also *p*- and *s*-consciousness), actual artificial agents (robots endowed with interactivity, autonomy and adaptability), and logically possible zombies (agents almost like humans, but for the fact that they lack *p*- and *s*-consciousness). To make things interesting, zombies will (at least appear to) be "switched-on" and able to exchange first-order information about the environment and interact with it as cognitively effectively, on average, as any ordinary conscious agent. The goal will be to establish not whether one belongs to the human type of agents or whether one knows that one does, but how one knows that one does. This will involve devising four versions of the knowledge game. The first two versions can be won by all inferential agents. This guarantees initial fairness and avoids begging the question. However, a third, more difficult version can be won only by non-artificial agents like us and the zombies. And a final, fourth version can be won only by *s-conscious* agents

---

theorem prover that uses various versions of the theorem as benchmarks (http://www.lwb.unibe.ch/index.html).

[5] For an approach to Dretske's question in terms of self-awareness see Werning [forthcoming]. Lycan [2003] argues that the inner sense theory can be defended against Dretske's criticism.

like us. I shall then argue that the presence of *s-consciousness* implies the presence of *p-consciousness*. So, if you win all versions, first, you are neither an artificial agent nor a zombie but a *p-* and *s-conscious* agent; second, you now know (assuming you did not already) that you are not a zombie; and third, you also have a way of explaining how you know that you are not a zombie, by pointing to you victory of the knowledge game. And since the test is only a sufficient but not a necessary condition to qualify as not a zombie, nothing is lost if one does not pass it. After all, you may still be a good driver even if you do not pass a driving test.

**3. The first and classic version of the knowledge game: externally inferable states**

A guard challenges three prisoners *A, B* and *C*. He shows them five fezzes, three red and two blue, blindfolds them and makes each of them wear a red fez, thus minimising the amount of information provided. He then hides the remaining fezzes from sight. When the blindfolds are removed, each prisoner can see only the other prisoners' fezzes. At this point, the guard says: "If you tell me the colour of your fez you will be free. But if you make a mistake or cheat you will be executed".

The guard interrogates *A* first. *A* checks *B*'s and *C*'s fezzes and declares that he does not know the colour of his fez. The guard then asks *B*. *B* has heard *A*, checks *A*'s and *C*'s fezzes, but he too must admit he does not know. Finally, the guard asks *C*. *C* has heard both *A* and *B* and immediately answers: "My fez is red". *C* is correct and the guard sets him free. As Dretske would put it: *C* is indeed in the state in which he says he is, and *C* knows that he is in that state or he would not have said so, the question is, *how does he know it?*

Take the Cartesian product of the two sets of fezzes. If there were three fezzes of each colour, we would have the following Table 1 (1 = red, and 0 = blue):

| Table 1 | a | b | c | d | e | f | g | h |
|---|---|---|---|---|---|---|---|---|
| *A* | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| *B* | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| *C* | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

Figure 1 The setting of the first version of the knowledge game

The prisoners know that they all know that there are only two blue fezzes, so ¬ h is *common knowledge*. This is a crucial piece of external information, without which no useful reasoning would be possible. Consider now *A*'s reasoning. *A* knows that, if *B* and *C* are both wearing blue fezzes, he must be wearing a red one (situation d). However, *A* says that he does not know, so now ¬ d is also common knowledge. *B* knows that if both *A* and *C* are wearing blue fezzes, he must be wearing a red one (situation f). However, *B* too says that he does not know, so *C* also knows that ¬ f. Moreover, since *B* knows that ¬ d, he also knows that, if he sees *A* wearing a red fez and *C* wearing a blue one, then he can only have a red fez (situation b). Since *B* says that he does not know, *C* also knows that ¬ b. Updating Table 1, the final Table 2, available to *C*, is (the top row indicates who knows which situation):

| Table 2 | | BC | | ABC | | BC | | ABC |
|---|---|---|---|---|---|---|---|---|
| | a | ¬ b | c | ¬ d | e | ¬ f | g | ¬ h |
| *A* | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| *B* | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| *C* | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

Figure 2 Who knows what at the end of the first version of the knowledge game

At this point, the game is over, since in all remaining situations {a, c, e, g} *C* is wearing a red fez. Note that *C* does not need to *see A* and *B*, so *C* could be blind. So, in a slightly different version, the prisoners are in a queue facing a wall, with *A* seeing *B* and *C*, *B* seeing *C*, and *C* looking at the wall. Despite appearances, the better off is still *C*.

### 3.1. A fairer version of the game

Sometimes the agents have a letter attached to their back or a muddy forehead, or play a card game (Fagin et al. [1995]). We only used 1s and 0s. The details are irrelevant, provided we are considering *externally inferable* states. Given an agent Ag, a state *S* and an environment *E*, *S* is an externally inferable state if and only if

i) in *E*, Ag is in *S*; and

ii) logical and informational resources can be polarised in such a way that

ii.a) only logical resources are in Ag;

ii.b) all informational resources are in *E*; and

ii.c) if Ag has access to *E*, then Ag can infer Ag's state *S* from *E*.

The prisoners exploit three environmental resources:

a) the nature and number of available states;

b) the observable states of the other prisoners;[6]

c) the other prisoners' answers;

plus the fact that they have common knowledge of (a)-(c).

Resource (c) is the only one that increases in the course of the game. This is unfair, for prisoner *A* can take no advantage of (c), whereas prisoner *C* cashes in on all the previous answers. Prisoner *B* is the most frustrated. He knows that, given his answer, if he were *C* he would be able to infer his (*C*'s) state. The fact that *B* cannot answer correctly before *C*, despite knowing that his answer will allow *C* to answer correctly without (*C*) even looking at *A* and *B* (recall that *C* may be blind), shows that *B* knows what it is like to be *C* – both in the sense of being the agent whose turn it is to answer the question, and in the sense of being in *C*'s given state – but that *C* is still "another mind" to *B*. *B* knows Table 2 as well as *C*, but cannot put this information to any use because he is not *C*. If there were no "other minds", there would be no difference in the location of *B* and *C* in the logical space of the knowledge game, but there is such a difference, so *B* and *C* are different, and *B* knows it. This, by the way, shows one way in which *B* can prove the existence of other minds.

To give a chance to every prisoner, the guard must interrogate all of them synchronically. *Mutatis mutandis*, the fair challenge goes like this:

*Guard*: "Do you know the colour of your fez?"

*A, B and C together*: "No".

Now all agents are in the state in which *B* was in the unfair game.

*Guard*: "Think again. Do you know the colour of your fez?"

*A, B and C together*: "No".

[6] The whole point of having a distributed system is that the components can communicate about their states. However, in our case this cannot be done by explicit acknowledgement of one's state, since the experiment relies on the agents not knowing already in which states they are. Therefore, the communication must be in terms of external observation, which requires some form of access. All this is easily modelled in terms of observable states, but it does not have to be. For example, the prisoners could be blindfolded and made to choose the fez to wear, one after the other, in such a way that each would know only which fez the other two have chosen. In this case, they would have to rely on their memories of observable processes. Note, finally, that in our version the communication is verbal and explicit, but in another version the prisoners are synchronised and merely asked to walk silently towards the door of the cell as soon as they know the answer. They all walk together after a given time.

Now they are all in the state in which *C* was in the unfair game, so they can immediately add, without being asked a third time:

*A, B and C together*: "Yes, it is red".

In the fair challenge, the prisoners work synchronically, no longer sequentially, and in parallel, as a multiagent system. The system is entirely distributed. It still relies on shared memory, but there is no centralised decision-taker, planner or manager, no CPU or *homunculus* that collects, stores and processes information and organises the interactions between the components. An interesting consequence is a net increase in efficiency. The multiagent system can now take full advantage of the resources (a)-(c) and extract more information from the environment, in fact all the available information, by excluding more alternatives under more constraints. Hearing *C*'s reply in the unfair challenge we only know that he is correct, but we do not know what *A* and *B* are wearing. Hearing the system, on the contrary, we come to know that all prisoners wear a red fez.

### 3.2. Winners of the classic version

Since all informational load is outsourced in the environment and *A*, *B* and *C* are *tabulae rasae* that behave like mere inferential engines, they are replaceable by artificial agents or zombies. Three Turing machines in a network could ascertain, inferentially, that they are all switched on (e.g. three green LEDs) and not off (e.g. two red LEDs). Put them in a black box, query the box about its own state, and you will obtain a correct answer. Having hidden some details, it seems that the box is magically "aware" of its own "switched-on" state. Clearly, the level of abstraction at which we observe the agent(s) makes a significant difference. Indeed, the solution of the prisoners' problem can be transformed into a computable algorithm generalisable to any finite number of interacting computer systems, something that turns out to be quite useful in industry.[7]

In one more variant of the classic version of the knowledge game, we can imagine that a robot, a zombie and a human prisoner team up to win the game. Externally inferable states do not allow us to discriminate between types of inferential agents.

[7] See for example DESIRE, a computational framework for DEsign and Specification of Interacting REasoning components to model distributed air traffic control that is based on the classic version of the knowledge game (Langevelde et al. [1992]; Brazier et al. [1994]).

To differentiate between the types of agents, we need a tougher game, and I shall suggest three alternatives. First, we can make the agents rely on some information made available by their newly acquired states themselves. This game will be analysed in the next section, where we shall see that Dretske is right: in this case too we cannot discriminate among different types of agents, although for different reasons from those just discussed. At this stage, not only do the agents not know how they know that they are not zombies, they might even doubt whether they are zombies or robots. We can then make the agents exploit whatever information is provided by the question itself. Artificial agents lose this version of the game. Finally, we can make the agent exploit the information implicit in their own answers. This version can be won only by conscious agents. It shows that whoever passes the test is not a zombie, that one can know this by winning the game, and that one can explain how one knows that one is not a zombie by referring to the way one wins the game. The last version of the game thus provides an answer to Dretske's question.

**4. The second version of the knowledge game**

The prisoners are shown five pairs of boots, all identical but for the fact that the three worn by the prisoners are torturing instruments that crush the feet, while the remaining two are ordinary boots. The guard plays the fair version of the challenge. Of course, the three prisoners answer correctly at once. Fezzes have only useless tassels, but torturing boots can *bootstrap*.

Bootstrapping is a technique that uses the input of a short sequence of instructions to make a system receptive to a larger set of instructions. Here one can slightly adapt the term to describe the new state of wearing torturing boots because

i) wearing them provides more external information (the short sequence) than wearing fezzes, namely a mechanical pressure;

ii) the prisoners are agents capable of receiving this extra information; and

iii) the extra information is sufficient to *verify* the state (the larger set of instructions) without having to *derive* it through a sequential or parallel process.

In bootstrapping states, the information about the "large" states becomes inferable through the interaction between the "short" state, its carrier and its receiver

(Barwise and Seligman [1997]).[8] Unfortunately, one cannot reach a satisfactory taxonomy of agents on these grounds.

Bootstrapping states are useless for discriminating between humans and zombies because the inference requires no *p-* or *s-consciousness* but only some sort of registration of the bootstrapping state as a premise to the successful inference. Since the underlying conscious life is the only difference between humans and zombies, whatever state is bootstrapping for the former may be assumed to be so (at least functionally) also for the latter and vice versa. But bootstrapping states are also useless for discriminating between artificial and non-artificial agents. A state is bootstrapping only *relationally*, depending on the source *and* the receiver of the information that indicates the state. So other types of agents can have their own types of bootstrapping states, which may or may not be bootstrapping for human agents (imagine the boots bear a barcode label). Since not all types of bootstrapping states are necessarily so for any human-like agent, a general bootstrapping game does not allow the necessary distinction between being able to play the game and winning it. For either the three types of agents are assessed on the basis of the same (types of) bootstrapping states accessible to all of them, in which case the game is useless, for they all win (participating is winning). Or the agents are assessed on the basis of different, i.e., their own, idiosyncratic (types of) bootstrapping states, accessible to only some of them, in which case the game is still useless, since each type of agent wins the game in which its own idiosyncratic states are in question (again, participating is winning).

One may object that precisely because some specific types of agents can be nomically associated to some types of bootstrapping states, the game can be modified so that the chosen (types of) bootstrapping states allow one to discriminate at least between artificial and non-artificial agents. But biological chauvinism won't help, as already shown by Turing [1950] and further argued by Dretske [2003], even if only for the distinction between zombies and humans. Selecting some specific (types of) feelings or experiences or perceptions to show that we and the zombies can perceive them as bootstrapping, but artificial agents do not, would be like choosing "heart-

[8] The *verification* process of a bootstrapping state requires more information (the "short" state) but less time (number of logical steps) than the *derivation* of the same state through external inference (imagine the case in which the torturing boots are also red and the prisoners are sitting at three tables and cannot see their own boots). Thus, verification capacities confer a selective advantage on the agent displaying them.

beating" as a criterion of discrimination. First, we are back to the "participating = winning" situation, this time in the converse sense that losing the game is equivalent to being unable to play it, not to playing it unsuccessfully. This makes the game not only unfair but above all uninteresting, for it is trivial to show that agents with access to different resources perform differently. Second, the game either presupposes the difference between types of agents that it is supposed to identify, thus begging the question, or it misses a fundamental point, the indiscernibility of the differences between the bootstrapping experiences. Zombies are almost like us: they z-feel the z-pain of the bootstrap, and they z-verify the corresponding z-state in ways that are either identical to ours or at least not discernible (for us) from them anyway. Dretske draws roughly the same conclusion when discussing zombies' *protopain*. Likewise, it would be very simple to engineer artificial agents capable of a-feeling the pressure of the "painful" boot or any other bootstrapping state we may choose. In either case, as far as we know, no difference between experiencing, z-experiencing and a-experiencing torturing boots can be usefully exploited within the game. So Dretske is right. Appeal to self-booting experiences won't do. You cannot answer the "how" question by relying on them. We need a different version of the game.

**5. The third version of the knowledge game**

So far the players have taken advantage of (their common knowledge of) the information provided by (i) the nature and number of assignable states; (ii) the observable states of the other agents; (iii) the other agents' answers; and (iv) the assigned states, when they are bootstrapping. A source that has not yet been exploited is the question itself.

Suppose the prisoners are offered five glasses. They are informed that three contain water and two contain a totally-deafening beverage. We play the game *unfairly*, i.e., sequentially, not synchronically. The first thing prisoner *A* hears is the guard shouting his question. Of course, the prisoner answers correctly at once. To him and to us, the question is trivially self-answering. Yet, why it is so is much less obvious. The guard's question (*Q*)

i) does not *entail* the answer;

ii) does not *implicate* (in Grice's sense) its answer, for this is not a matter of implicit meaning of *Q* or intention of the utter of *Q*;

iii) does not *presuppose*[9] its answer, for *Q* cannot be answered by relying on the background information that the speaker and the listener tacitly assume (Groenendijk and Stokhof [1994], § 6.4.5);
iv) is not *self-fulfilling*, that is, it does not contain its answer in the "fridge-paradox" sense that asking *Q* is equivalent to implementing whatever state *S* is whose existence *Q* is about (like opening the fridge to check whether the light is on);
v) is not *loaded*, for *Q* is loaded only if the respondent is not committed to (some part of) the presupposition of *Q* (Walton [1991], 340).

Self-answering questions are not the subject of much analysis in erotetic logic.[10] Perhaps they are too trivial. Sometimes they are even confused with rhetorical questions, which are really assertions under cover. Yet a self-answering question is not one that requires no answer, or for which the questioner intends to provide his or his own answer. It is a question that answers itself if one knows how to interpret it, and this can be achieved in several ways. The erotetic commitment of the question can be external. For example, asking a yes/no question while nodding may count as an externally, pragmatically self-answering question. Or the erotetic commitment can be internal. "How many were the four evangelists?" is an internally, semantically self-answering question. In our case, the erotetic commitment is neither internal nor external but relational. The question about the agent being in a certain state is self-answering in a more complex way, for the answer is *counterfactually embedded* in *Q* and it is so somewhat "indexically"[11] since, under different circumstances, the question or the questioning would give nothing away (henceforth this is what I shall mean by self-answering question).

For *A* to extract from the self-answering *Q* the information that *A* is in *S*, something like the following is required:

1. *A*, *B* and *C* can each be set in a new state, either *S* or ¬ *S*
2. *A* receives the information contained in (1)
3. *A* is set in a new state, either *S* or ¬ *S*

[9] "A *presupposition of a question is* defined as a proposition that the respondent becomes committed to in giving any direct answer to the question" (Walton [1991], 338).
[10] In their excellent survey, Groenendijk and Stokhof [1994] pay no attention to self-answering questions.
[11] During the conference in Glasgow, Selmer Bringsjord and Patrick Grim pointed out to me that this use of "indexically" may not be entirely appropriate and could generate confusion. I appreciate their concern, but I do not know of any other term that would express the point equally well. The reader is warned that the qualification is slightly unorthodox.

4. *A* receives the information contained in (3)
5. *A*'s new state is *S*
6. *A* does not receive the information contained in (5)
7. *A* receives the question *Q* about the nature of *A*'s new state
8. *A* receives the information contained in (7)
9. *A* reasons that if *A* were in ¬ *S* then *A* would be in some state *D*; but if *A* were in *D* then *A* could not have received *Q*; but *A* received *Q*, so *A* could receive *Q*, so *A* is not in *D*, so *A* is not in ¬ *S*, but *A* is in either *S* or ¬ *S*, so *A* is in *S*.
10. *A* answers that *A* is in *S*.

An interesting example of this new version of the knowledge game is provided by Hobbes and Gassendi. At different stages, they both object to Descartes that states such as "walking" or "jumping" may replace "thinking" within the Cartesian project. "Ambulo ergo sum" would do equally well, they argue. However, Descartes correctly replies that they are both mistaken (and they were). "Are you (am I) thinking?" is self-answering in the sense just defined, but "are you (am I) walking?" is not. As we shall see in the next game, a zombie can jump and walk but he still cannot infer (let alone be certain) from this that he exists, for *he* does not (indeed cannot) know that *he himself* is jumping and walking. Whereas, even if we perceive ourselves jumping and walking, we may still wonder whether we are dreaming, in which case it is the activity of wondering (in other words: thinking) that one may be dreaming that makes the difference, not the dreamt state itself; or we may wonder whether we are zombies, and if so, whether we are zombies dreaming that they are walking, in which case too there is still nothing intrinsic to the jumping or to the walking that will enable us to tell the difference, i.e., to answer Dretske's "how" question.

*Extracting* (as opposed to *verifying* or *deriving*) information (the erotetic commitment) about states from self-answering questions about those very states requires agents endowed with advanced semantic capacities. These are often clustered under broader and more general labels such as intellect, reason, intelligence, understanding, high-order cognition or mind. In order to be less inclusive and to stress their procedural nature, I suggest we opt for *reflection*.

Reflection is only a useful label and it is not to be understood here as referring to some higher-order awareness or cognition, if lower-order awareness or cognition is a sense of self, or consciousness. So far we have avoided relying on consciousness (in any sense of the word) and we should keep resisting the temptation to cheat or to beg

the question. Reflection is not meant to refer to privileged access, introspection or psychological awareness of internal states either, no matter of what kind and order. Rather, it is to be understood as a label for the semantic capacity of transcendental (backward) inference, from understanding the question to understanding its conditions of possibility and hence its answer.

Reflection so understood is something that artificial agents do not enjoy yet. The reader may be acquainted with "reflective" artificial agents that can win the classic knowledge game (Brazier and Treur [1999]), but that description is only evocative. Architectures or programs for computational systems (of AI) and systems for machine learning are technically called "reflective" when they contain an accessible representation of themselves that can be used (by themselves) e.g. to monitor and improve their performance. This seems to be what Ned Block [1995] has termed, in human agents, "access-consciousness" (see Bringsjord [1997]). However, what is known as *reflective computing* is only a case of metaprogramming,[12] and the knowledge game does not contradict this deflationist view, since, strictly speaking, the axiomatization of the reasoning involved requires only standard first-order logic (McCarthy [1971-1987] and McCarthy [1990]) without any appeal to introspection, even if it can be analysed using e.g. epistemic logic or the BDI (Belief, Desire, Intention) architecture (Rao and Georgeff [1991]).

Current artificial agents are unable to take advantage of the self-answering nature of the question because they are intellectually and semantically impaired, somewhat like Descartes' animal automata or Spinoza "sceptical" automata. Reflection is an AI-complete problem, i.e., a problem whose solution presupposes a solution to the "strong AI problem", the synthesis of a decent (to many this is synonymous with human) level of commonsensical intelligence endowed with some semantic skills. As we still lack anything even vaguely resembling a semantically proficient and intelligent artificial agent, this version of the knowledge game suffices to discriminate between them (the artificial) and us (zombies and humans). Let me now qualify this claim.

First, to be answered correctly, a self-answering question requires both understanding of the content of, and a detachment from, the question itself. Self-answering questions are part of the *frame problem*. A normal query works like an

[12] Barklund [1995] and Costantini [2002] are two valuable surveys with further references to the "three wise men" problem.

instruction that pushes an agent Ag into a search-space, where a correct symbolic manipulation can identify the state of the agent itself. But a self-answering question pulls a reflective agent in the opposite direction, in search of what his own state must be if the question is receivable/accessible and intelligible in the first place. Now, some counterfactual truths concerning a variety of type-situations can be pre-programmed (soft-encoded, hard-wired or "interfaked", i.e. faked through an interface) in our artificial agents. As in the Turing test, this is trivially achievable, yet it is not a solution but only an *ad hoc* and brittle trick. It is the difference between winning a type of game or just a specific token of it. Contrary to artificial agents, zombies and humans can be assumed to have a full and intelligent command of language, and hence enjoy a counterfactual-reflective understanding of the semantics of an open-ended number of indexical questions. If we repeat the test, in principle, if the question is self-answering, zombies and humans should be able to appreciate it, but artificial agents cannot. Any digital make-up is here only a boring "catch me if you can" of no conceptual interest.

Second, I specified above that we have tested the difference between artificial agents and zombies by making them play the third version of the knowledge game *unfairly*, i.e. sequentially not synchronically. Zombie *A* passes the test, whereas robot *A* does not. What happens if we let the artificial agents play the game *fairly*? A parallel, multiagent system of artificial agents would still fail the test. This because the answers provided by the other machines are non-informative (they do no provide further, informational constraints on *C*'s options), so they do not really improve the overall performance. The last artificial agent *C*, or the whole system working in parallel, is in the same predicament as the first artificial agent *A*, or the system working sequentially. So the performance of a single zombie will still be different from the performance of a system of artificial agents. Zombie-like behaviour (or a semantic engine) cannot yet been obtained by coordinating several artificial agents (syntactic engines). If it could, we would have found a way of building semantic systems.

Third, my previous claim comes with a "best before" date: *current* and *foreseeable* artificial agents as we know them cannot answer self-answering questions, either in a stand-alone or in a multiagent setting. What is *logically possible* for (or achievable at some distant time in the future by) a single artificial agent or for an artificial multiagent system is not in question. I explained at the outset that we are

not assuming some science fiction scenario. "Never" is a long time, and I would not like to commit myself to any statement like "artificial agents will never be able (or are in principle unable) to answer self-answering questions". The knowledge game cannot be used to argue that AI or AC (artificial consciousness) is impossible in principle. In particular, its present format cannot be used to answer the question "How do you know you are not a *futuristic*, intelligent and conscious artificial agent of the kind envisaged in *Blade Runner* or *Natural City*?". As far as artificial agents are concerned, the knowledge game is a test to check whether AI and AC have been achieved. We shall see that, given the difference between us and zombies, as Bringsjord [1999] has rightly noticed, if one day artificial agents will pass the test, we shall have created zombies.[13]

Unlike artificial agents, zombies are reflective in the sense specified above, for (the starting hypothesis is that) they share with us everything but *p*- and *s*-*consciousness*. So Descartes, Spinoza, Leibniz [1995] and Dretske are right: we and they can win this version of the knowledge game and nobody could spot the difference. Is there any other source of information that a conscious agent can exploit inferentially but a zombie cannot?

Recall that the difference is supposed to rest on the subjectively conscious nature of the states in the (*p*) and (*s*) senses. This is mirrored in the nature of the corresponding reports. A zombie *Z* knows that *Z* is in state *S*, but does not know that *he* is *Z*, or that *S* is *his* state, nor does *Z* know what it is like for *himself* to be in *S*. A human agent *H*, on the contrary, will find it difficult to dissociate himself from his own states, which are always "proprietary". To *H*, *H*'s states are first of all his own states or states as he experiences them (thus-states), or of which he is conscious, at least in so far as his attention can be called upon *S*. A detached (third-person or zombie-like) perspective on one's own thus-states can be acquired, but it is not the default option and, if adopted, would seem rather contrived.

[13] Having made this much clear, I entirely agree with Searle [1992] and Bringsjord [1999] in their criticism of computationalism. As Bringsjord writes in his defence of Searle's position against Dennett, in current, computational AI "[…] "the person building project" will inevitably fail, but […] it *will* manage to produce artifacts capable of excelling in the famous Turing Test, and in its more stringent relatives. What sort of artifacts will these creatures be? I offer an unflattering one-word response: Pollock, Dennett, and like-minded researchers are busy building… *zombies*". Indeed, I am even more pessimistic than Bringsjord. I do wonder, however, whether neo-Frankensteinian AI may be computer-ethically questionable (Floridi and Sanders [2004]).

This intuition can be made profitable by exploiting a last source of information about the agent's state, namely his own answer, thus coming full circle (you may recall that in the first version each agent takes advantage of the other agents' answers, not yet of his own).

## 6. The fourth version of the knowledge game

The three prisoners are offered five tablets: three are completely innocuous and two make the agents totally dumb. As before, we play the game *sequentially*. For reasons that will be discussed in the following section, we shall focus only on prisoner *A*'s performance. *A* cannot know which tablet he has taken in terms of externally inferable, bootstrapping or self-answering states. As usual, all forms of privileged or direct accesses to his state are also excluded *ex hypothesi*. *A* hears the question and is then allowed to answer. Since he has no way of knowing or inferring whether he is in a dumb state, he answers by reporting his state of ignorance. Now, *whatever A* says to communicate his state of ignorance, e.g. "Heaven knows", [14] either

a) his verbal report about his state of ignorance triggers no further reaction; or

b) his verbal report about his state of ignorance triggers a counterfactual reasoning of the following kind: "had I taken the dumbing tablet I would not have been able to report orally my state of ignorance about my dumb/non-dumb state, but I have been and I know that I have been, as I have heard myself speaking and saw the guard reacting to my speaking, but this (my oral report) is possible only if I did not take the dumbing tablet, so now I know that I am in a non-dumb state, hence I know that I have not taken the dumbing tablet, and I know that I know all this, that is, I know that my previous state of ignorance has now been erased, so I can revise my statement and reply, correctly, in which state I am, which is a state of not having taken the dumbing tablet, of knowing that I haven't, and – by going through this whole process and

[14] It would be more elegant to allow the agents to reply in the usual way, by saying "I do not know" when they wish to acknowledge their state of ignorance, but during the discussion of this paper I noticed that several people found this confusing. For someone might wonder whether passing the test is somehow connected to the capacity of issuing first-person report, and this is definitely not the point. Nothing hangs on the specific verbal capacities of the agents or on the language employed. The game would work equally well if the tables were "arms-paralysing" and the agents could reply only by raising their arms. Basically, any form of communication will do, as long as one can take advantage of the detachment between reporting and reported agent by identifying a state *S* such that the reporting agent *A* knows that he is in *S* only if *A* knows that he himself is the reporting agent *x* in his own description of his lack of information about *x*'s state.

passing the test – of knowing how I know both that I haven't and that I know that I haven't".

In case (a), *A* does not correct himself and so he fails the test: he does not know in which state he is. Literally. The agents who fail the test do not know that it is themselves that they are hearing talking. If they did, they would pass the test. These agents lack *s-consciousness*.

In case (b), *A* corrects himself, so he is able to pass the test by saying in which state he is. In order to be able to infer his actual state on the basis of his initial report about his lack of information about their state, the player needs to be able to identify

1. the agent reporting (e.g. by uttering an interjection to communicate his lack of information) that
2. the agent is in a state of ignorance about
3. the agent's empirical state in question (dumbness/non-dumbness, in our example)

as the *same* agent, and then this agent as himself, that is, as the agent *A* who is playing the knowledge game and to whom the guard addressed the question concerning his state after having taken the tablet. But this means that players can pass the test only if they realise that it is themselves and their own states that they are talking about. Winners of this final version of the knowledge game need to be *s-conscious* agents, so they cannot be zombies.

We have seen that a difference between artificial agents and zombies/humans is that the latter are capable of *counterfactual reflection*, something that requires *semantics*. A difference between zombies and humans is that the latter are also capable of *subjective reflection*, something that requires not only *semantics* but also *consciousness*. The time has come to use the knowledge game to answer Dretske's question.

**7. Dretske's question and the knowledge game**

How do you know you are not a zombie? The briefest answer is: by playing as *A* and winning the last version of the knowledge game. The long answer will take some clarifications.

By winning the fourth version of the knowledge game, the agent *A* (i) shows that he is not a zombie, (ii) experiences what it is like to be *s-conscious* (and hence not a zombie in this sense) by identifying himself as himself while being in *S*; and experiences what it is like to be *p-conscious* of *S* (and hence not a zombie in this

sense too) by realising what it is like to be in a certain state *S* and be able to know and explain that he is in that state *S*. Shouldn't one, having won the game, wonder whether what one experienced as state *S* might also be z-experienced by a zombie? No, because, contrary to what happens in the third version of the game, what is in question now is not the mere experience of being in a state *S*, nor the knowledge of being in that state, nor just the simple *p-consciousness* of being in *S*, but the *p-consciousness* of being in *S* brought about by (what is the other side of consciousness namely) the *s-consciousness* of being the agent who is in *S*, and this twofold consciousness brings with itself the further knowledge that this is as good evidence as anything that one is not (indeed cannot) z-experiencing *S*, for if one did, then one would not be able to win the game as one did.

Perhaps a different approach, based on the logical order of the questions, might be useful to clarify the argument. Consider the following statements made of question + answer:

1) are you a not a zombie? yes;

2.1) do you know that (1)? yes;

2.2) how do you know that (1)?

3) do you know that (2.1)? yes.

Points (1) and (2.1) are conceded. Point (2.2) is not an implicit challenge of (1) or (2.1) but it is Dretske's question. We have seen that we cannot use the knowledge game to answer (2.2) by relying on (1) and (2.1). The fourth version of the game shows that (2.2) can be answered by appealing to a higher level, namely (3), what in epistemic logic would go under the general label of the KK thesis. Should we not repeat (2.2) at an even higher level, and wonder how one knows that one knows that (1)? No, because the escalation from (2.1) to (3) brings about the so-called "common knowledge" phenomenon: once an agent knows that he knows that he is not a zombie, then he also knows that he knows that… he knows that he is not a zombie. So how do you know that you are not a zombie? By passing a test that proves that you know that you know that you are not, which implies that you know that you are not, which implies that you are not; and then by pointing to this whole process, and the corresponding test that brings it about and the common knowledge it generates, to explain how you know all this.

Two final clarifications can now be introduced as answers to two main objections.

First, we have seen that Dretske's question concerns primarily knowing that one is *p-conscious*. However, one may contend that winning the last version of the knowledge game indicates, at best, that the agents in question are *s-conscious*. If so, it says nothing about *p-consciousness* and *a fortiori* nothing about how one knows that one is *p-conscious*.

Second, one may argue that, if we allow the agents to play the fourth version of the game until *C* answers, or *fairly*, i.e. *synchronically*, it seems that *C* (or the multiagent zombie system) will pass the test, and this shows that the test itself is not a good test anyway.

Both objections are reasonable but easily answerable.

Consider the first objection. The knowledge game shows not only that an agent *A*, who is *s-conscious*, can tell in which empirical state *S* he is (e.g., he is not dumb), but also that, by going through the process of identifying his correct empirical state *S* while being in *S*, the agent *A* comes to realise what it is like to be in that state *S*, that is, he is also able to appreciate the subjective nature of the state *S* involved, at least insofar as this triggers the consciousness of being in *S*. Stripped of its qualitative nature, the reasoning is that, if the agent enjoys no *p-consciousness* of any kind then he enjoys no *s-consciousness* of any kind, but winning the knowledge game shows that the agent enjoys *s-consciousness* of some kind, so it also shows that he enjoys some *p-consciousness* of some kind (the "kind" is contingent on the setting of the game and the choice of the state *S*).

Someone may wish to resist the previous line of reasoning by arguing that the form of the *modus tollens* is hardly questionable, but that the implication to which it is applied is controversial. The reply is that there are plenty of reasons for being confident about the implication as well.

To begin with, *p-* and *s-consciousness* seem to be strictly related. Indeed, they may be related by a double implication, the tenability of one half of which has already been established by Kriegel [2004]. Kriegel has argued, convincingly, that some forms of *p-consciousness* imply some forms of *s-consciousness*. If one does not find this compelling, the ball is in his court. Suppose we therefore accept that some forms of *p-consciousness* imply some forms of *s-consciousness*. It then seems much less

difficult to argue for the other half of the equation,[15] from some forms of *s-consciousness* to some forms of *p-consciousness*, which is the implication needed above, in the negative form that absence of the latter implies absence of the former. If an agent has some (second- or higher-order) sense, or is (introspectively) aware of, his personal identity (including his knowledge that he thinks) and (first- or lower-order) experiences, both mental and perceptual (including his knowledge of what he is thinking), then he must also have some experience of the qualitative, subjective, personal or phenomenological properties of the state in which he is. You cannot be *x*, be *s-conscious* that you are *x*, know that you are *s-conscious* that you are *x* and yet not know (i.e. be *p-unconscious* of) what it's like to be *x*, because the knowing (being *p-conscious* of) of what it's like to be *x* is just equivalent to knowing (being *s-conscious* of) what it's like to be yourself.

Our skeptical interlocutor may still be unconvinced. Appealing to some common sense; showing that in some cases half of the double implication is already proved and the other half is even more reasonable; shifting the burden of proof; reasoning that you cannot be *s-conscious* in a vacuum of *p-consciousness* of something; all this may still appear to be insufficient. Luckily, we can leave the skeptic to his own doubts because Dretske himself seems happy to concede, following several other philosophers, that indeed *p-consciousness* is inferably related to *s-consciousness*. The point at stake is not that Dretske believes that *p-* and *s-consciousness* are unrelated, but that he suggests that relating *p-consciousness* to *s-consciousness* makes no difference to whether his "how" question can be answered. Yet, the knowledge game shows that this conclusion is unjustified. It does make a difference because by then having a test based on the presence of *s-consciousness while* having a *p-consciousness* of a particular state, you can have the experience of what it is like to be *p-conscious* of *S and* have a way of showing how you know that you are *p-conscious* of *S*. Dretske was right when in the past he thought the argument based on *s-consciousness* was convincing.

So, going back to our skeptic, the worst scenario is that at least Dretske and I share the same presupposition: *p-* and *s-consciousness* are two sides of the same coin: take it all or leave it all. Disjoining the two is misleading. Since the two are

[15] Kriegel [2003] comes very close to arguing for this second half of the double implication when he supports the thesis that "a mental state is conscious when, and only when, it involves implicit self-awareness". I'm grateful to Kriegel for having called my attention to this article.

correlated, showing how you know that you are *s-conscious* while you are in *S* is at the same time a way of showing how you know you are also *p-conscious* of *S*.

Consider now the second objection. If we let the agents play the game until the end (*C*'s answer) or synchronically (as a multiagent group), the answers by the three zombies (e.g. three "Heaven knows") will provide increasing informational constraints on *C*'s (or on the multiagent zombie system's) options. The zombies already have the information that there are only three players. Therefore, from hearing first one, then two and finally three voices, it will be possible to infer that all players have taken innocuous tables, and this without recurring to any sort of consciousness. Eventually, *C* (or the multiagent zombie system), on hearing three voices, will infer that the three zombies have taken the innocuous tablets, he will correct his statement and pass the test. So can the test be passed by *C* and hence by a multiagent zombie system? The right answer is: it should but it does not matter.

Of course, each zombie still does not know that he himself has taken an innocuous tablet because otherwise he would not have been able to answer in the first place, but this is not relevant here. What are relevant are two other considerations.

On the one hand, the test devised is fair only if your performance, as a whole conscious agent, can in principle become indistinguishable from the performance of yourself considered as a system each part of which (everybody should agree) is not conscious. Now zombies may just be such parts: we have assumed that each of them has all the (perceptual, semantic, cognitive, logical and so forth) capacities that you have, but for the fact of being *p-* and *s-unconscious*. So, if they coordinate their efforts, apart from a lot of spare capacity and redundancy (recall that each zombie is like a conscious-less you), we should expect the performance of a system of zombies to be as successful as yours, *given the constraints offered by the knowledge game*. In other words, you and the multiagent zombie system should appear to an external observer as performing equally well: you should both win it in two steps. This guarantees fairness and some plausibility, but it does not mean that the multiagent zombie system is conscious. For, on the other hand, recall that, precisely because you and a system of zombies would be able to win the fourth version, we have compared your performance as agent *A* to that of a single zombie in the same position, not to the whole system of interacting zombies. And because *A* does not yet have all the constrains that will be available to *C*, you can pass the test but a zombie won't, nor will a system of zombies, insofar as it would have to be considered in its turn as a

single agent *A* in need to coordinate its answers with other two multiagent zombie systems *B* and *C*.

A general lesson to be learnt from the previous discussion is that, when discussing how agents could win the third version, we saw that we cannot obtain zombie-like performance by coordinating the behaviour of our current conscious-less artificial agents. Whereas, we now know that, as far as winning the fourth version of the game is concerned, we could obtain close-to-conscious performance if only we could create multiagent zombie systems.

After having asked "how do you know you are not a zombie?", Dretske comments "Everything you are aware of would be the same if you were a zombie. In having perceptual experience, then, nothing distinguishes your world, the world you experience, from a zombie's." I agree. This is the hypothesis supporting the thought experiment. Dretske then asks: "This being so, what is it about this world that tells you that, unlike a zombie, you experience it?". He answers "nothing", and I agree, at least insofar as the first version of the knowledge game shows that externally inferable states are useless in this respect. Dretske finally asks: "What is it you are aware of that indicates you are aware of it?". His and my answer is again "nothing". I have argued in favour of this point by showing that self-booting and reflective states are also useless. Dretske asks no more questions and concludes: "We are left, then, with our original question: How do you know you are not a zombie? Not everyone who is conscious knows they are. Not everyone who is not a zombie, knows they are not. Infants don't. Animals don't. You do. Where did you learn this? To insist that we know it despite there being *no identifiable way* (my emphasis) we know it is not very helpful. We can't do epistemology by stamping our feet. Skeptical suspicions are, I think, rightly aroused by this result. Maybe our conviction that we know, in a direct and authoritative way, that we are conscious is simply a confusion of what we are aware of with our awareness of it". I have argued that this pessimistic conclusion is premature because it does not take into account the agent's inferential interactions with other agents as discussed in the fourth version of the knowledge game.

**8. Conclusion: some consequences of the knowledge game**

In the knowledge game there are four sources of information: the environment, the states, the questions and the answers. Agents are assigned predetermined states using the least informative setting (the initial difference in states is not imported within the

system, whose components are assigned equal states). They are then assessed according to their capacities to obtain information about their own states from these sources inferentially. Depending on the chosen state, each source can identify a type of game and hence a class of players able to win it. The communicative and logical nature of the game excludes even very intelligent mammals from participation, including infants, chimpanzees, orang-utans and bottlenose dolphins, who consistently pass the test of mirror self-recognition (Allen [2003]). This, however, is not an objection, since the question addressed in this paper is not how you (a grown-up person who can fully understand the "how" question) know that you are an animal. Conversely, whatever the answer to the latter question is, it certainly cannot rely on some unique logical capacities you enjoy.

The knowledge game is an informative test and is not meant to provide a *definition* of intellectual or semantic abilities or of consciousness. It is not a defence of an inferential theory of consciousness either, nor does it provide ammunition for the displaced perception model. Like the Turing test for AI, it purports to offer something much weaker, namely a reliable and informative criterion to discriminate between types of (inferential) agents and hence a way to answer the "how" question, without relying on that foggy phenomenon that is human introspection.

The criterion employed is more than a successful means of identification because it is also informatively rich. An informatively poor criterion would be one that used an otherwise irrelevant property *P* to identify *x* successfully. To use an example à la Donnellan, at a party you may successfully identify Mary's friend as the only man in the room – this also conveys quite a bit of extra information about him – or as the person who is closer to the window, which is true at that precise moment but otherwise very poor informatively. The knowledge game relies on relevant and significant properties that characterise agents in an informatively rich way. This is like cataloguing animals according to their diets. The trap would be to reduce (or simply confuse) what matters most in the nature of *x* to (or with) what makes an informatively rich difference between *x*, which has *P*, and *y*, which lacks *P*. An agent Ag may have the unique capacity to infer his own states and yet this may not be the most important thing about Ag, *pace* Descartes. We avoid the trap if we recall that our task is to answer Dretske's question. Consciousness-centrism may be perfectly justified and even welcome, but it is a different thesis, which requires its own defence; it is not what the knowledge game is here to support.

The last version of the game suggests a view of consciousness as *subjective reflectivity*, a state in which the agent and the I merge and "see each other" as the same subject. It seems that artificial agents and zombies are entirely decoupled from the agents they can issue reports about but which they cannot identify as themselves. Animals, on the other hand, seem wholly coupled to external information. Humans are both decoupled from their environment and coupled to themselves, that is, they appear to constitute themselves as centres of subjective reflectivity, prompted by, but independent of, the environment. Consciousness is comparable to a mathematical fixed point: it occurs as a decoupling from reality and a collapsing of the referring agent and the referred agent. I suggest this is what lies behind the description of the acquisition of consciousness as a sort of "awakening".

This perspective has some surprising consequences. One of the most interesting concerns the transcendental nature of the I.[16] In the final knowledge game, states are inferentially appropriated by the agent as his own (*p-consciousness*) only because the agent is already conscious of himself as himself (*s-consciousness*). Once unpacked, this logical priority may mean that agents are *p-conscious* of their perceptual/internal contents not only *after* or *if* but also *because* they are *s-conscious* of themselves. It certainly means that it is not true that they are *s-conscious because* they are *p-conscious* (the *if*, we have seen, has been accepted following Kriegel [2004]). Perceptual or internal contents of which the agent is conscious do not carry with themselves inferentially the (information that the agent has) consciousness of the content itself or of himself as an extra bonus. Perhaps *s-consciousness* is not constructed from perceptual and internal knowledge bottom-up but cascades on it top-down. This IBM ("I Before the Mine") thesis is a strong reading of Searle's view that "the ontology of the mental is an irreducibly first-person ontology" (Searle [1992], 95). Adapting Harnad's phrase, zombies are empty homes but your home is wherever your self is.

If *s-consciousness* really has a logical primacy over the conscious-ed contents, I doubt whether the IBM thesis might be reconciled with some sort of naturalism. It is certainly not externalist-friendly, if by externalism one basically refers, ontologically, to a thesis about where the roots of consciousness are – outside the mind – rather than, heuristically, to a thesis about where the search for them can start. For the knowledge

[16] I am grateful to Susan Stuart fro having called my attention to this point. Of course this is not to say that she would agree on my interpretation of it.

game shows that, in explaining consciousness without relying on introspection, we still cannot progress very far by relying only on environmental information.

The knowledge game coheres much better with an internalist perspective, with an important proviso. Sometimes semantic and rational agents can obtain information about their own states only if they interact successfully in a collaborative context rather than as stand-alone individuals.[17] The external source is a "Platonic", maieutic device, which has an eliciting role that is functionally inverse to the one attributed to the malicious demon by Descartes.[18] We have seen that the knowledge game promotes an intersubjective conception of agenthood, moving in the same direction as Grice's Cooperative Principle and Davidson's Charity Principle, while favouring an internalist view of *s-consciousness*.

[17] This "social" point is emphasized in Moody [1994], see also the several contributions and discussions of Moody's position in Symposium [1995], especially Bringsjord's convincing analysis.
[18] The examiner/guard and the questioning father in the muddy children version have a crucial role, for they guarantee *common knowledge* among the players, see Fagin et al. [1995]. This external source is a sort of ghost outside the machine.

**Acknowledgements**

I discussed several drafts of this paper at many meetings. It was the topic of a series of lectures on the philosophy of information at the University of Lisbon and I am grateful to Olga Pombo for that opportunity. A shorter version was then given as the *Alan Turing Lecture in Computing and Philosophy* at the European *Computing and Philosophy* Conference ECAP 2003 (University of Glasgow) and I wish to thank Susan Stuart for the invitation. A further revised version was then the topic of an invited lecture to the *Oxford Society for Artificial Intelligence*, of a graduate seminar in the history and philosophy of science organised by the Philosophy Department of the University of Bari, and of a graduate *Seminario di Logica e Filosofia Analitica* organised by Daniele Giaretta at the Philosophy Department of Padua University, and to whom I owe a most fruitful discussion. I am grateful to the participants in these meetings for their helpful discussions. In particular, I would like to acknowledge the help, useful comments and criticisms by Andrea Bianchi, Selmer Bringsjord, Massimiliano Carrara, Daniele Giaretta, Gian Maria Greco, Patrick Grim, Uriah Kriegel, Paul Oldfield, Gianluca Paronitti, Claudio Pizzi, Jeff Sanders, Susan Stuart and Matteo Turilli. Fabrizio Floridi provided the specific example of the three fezzes. Kia Nobre read the final version of this paper and made me aware of several crucial implications. If there are still obvious mistakes after so much feedback, I am the only person that should be embarrassed by them.

**References**

Allen, C. 2003, "Animal Consciousness" in *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Alston, W. 1971, "Varieties of Privileged Access", *American Philosophical Quarterly*, 8, 223-241.

Alston, W. 1986, "Epistemic Circularity", *Philosophy and Phenomenological Research*, 47, 1-30.

Barklund, J. 1995, "Metaprogramming in Logic" in *Encyclopedia of Computer Science and Technology*, edited by A. Kent and J. G. Williams (New York: Marcel Dekker), vol. 33, 205-227.

Barwise, J. 1988, *The Situation in Logic* (Stanford, CA: Center for the Study of Language and Information).

Barwise, J., and Etchemendy, J. 1987, *The Liar: An Essay on Truth and Circularity* (New York; Oxford: Oxford University Press).

Barwise, J., and Seligman, J. 1997, *Information Flow: The Logic of Distributed Systems* (Cambridge: Cambridge University Press).

Block, N. 1995, "On a Confusion About a Function of Consciousness", *Behavioral and Brain Sciences*, 18, 227-247.

Brazier, F. M. T., and Treur, J. 1999, "Compositional Modelling of Reflective Agents", *International Journal of Human-Computer Studies*, 50, 407-431.

Brazier, F. M. T., Treur, J., Wijngaards, N. J. E., and Willems, M. 1995, "Formal Specification of Hierarchically (De)Composed Tasks", in *Proceedings of the 9th Banff Knowledge Acquisition for Knowledge-based Systems workshop, KAW'95*, Calgary, edited by B. R. Gaines and M. A. Musen (SRDG Publications, Department of Computer Science, University of Calgary), 25/21-25/20.

Bringsjord, S. 1997, "Consciousness by the Lights of Logic and Commonsense", *Behavioral and Brain Sciences*, 20, 144-146.

Bringsjord, S. 1999, "The Zombie Attack on the Computational Conception of Mind", *Philosophy and Phenomenological Research*, 59(1), 41-69.

Conway, J. H., and Guy, R. K. 1996, *The Book of Numbers* (New York: Copernicus).

Costantini, S. 2002, "Meta-Reasoning: A Survey" in *Computational Logic: Logic Programming and Beyond - Essays in Honour of Robert A. Kowalski*, edited by A. C. Kakas and F. Sadri (Springer-Verlag),

Ditmarsch, H. P. v. 2000, *Knowledge Games* (Amsterdam). University of Groningen, doctoral thesis in computer science, available online at http://www.ai.rug.nl/~hans/.

Dretske, F. 2003, "How Do You Know You Are Not a Zombie?" in *Privileged Access and First-Person Authority*, edited by B. Gertler (Burlington: Ashgate),

Elmer, J. 1995, "Blinded Me with Science: Motifs of Observation and Temporality in Lacan and Luhmann", *Cultural Critique*, 30, 101-136.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. 1995, *Reasoning About Knowledge* (Cambridge, Mass; London: MIT Press).

Floridi, L. 1996, *Scepticism and the Foundation of Epistemology: A Study in the Metalogical Fallacies* (Leiden: Brill).

Floridi, L., and Sanders, J. W. 2004, "The Method of Abstraction" in *Yearbook of the Artificial. Nature, Culture and Technology. Models in Contemporary Sciences*, edited by M. Negrotti (Bern: Peter Lang), Available online at http://www.wolfson.ox.ac.uk/~floridi/pdf/loa.ps.

Floridi, L., and Sanders, J. W. 2004, "On the Morality of Artificial Agents", *Minds and Machines*, 14(3), 349-379. Preprint available at http://www.wolfson.ox.ac.uk/~floridi/

Groenendijk, J., and Stokhof, M. 1994, "Questions" in *Handbook of Logic and Language*, edited by Van Benthem and Ter Meulen (North-Holland: Elsevier Science),

Groenendijk, J. A. G., Janssen, T. M. V., and Stokhof, M. J. B. (ed.) 1984, *Truth, Interpretation, and Information: Selected Papers from the Third Amsterdam Colloquium* (Dordrecht, Holland; Cinnaminson, U.S.A: Foris Publications).

Kriegel, U. 2003, "Consciousness as Sensory Quality and as Implicit Self-Awareness", *Phenomenology and the Cognitive Sciences*, 2(1), 1-26.

Kriegel, U. 2004, "Consciousness and Self-Consciousness", *The Monist*, 87, 185-209.

Lacan, J. 1988, "Logical Time and the Assertion of Anticipated Certainty", *Newsletter of the Freudian Field,* 2, 4-22. Originally written in March 1945, this was first published in *Écrits*, pp.197-213, 1966.

Langevelde, I. A. v., Philipsen, A. W., and Treur, J. 1992, "Formal Specification of Compositional Architectures", in *Proceedings of the 10th European Conference on AI, ECAI-92*, edited by B. Neumann (John Wiley & Sons), 272-276.

Leibniz, G. W. 1995, "Monadology" in *Philosophical Writings*, edited by G. H. R. Parkinson (London: Dent; first published in Everyman's Library in 1934; published, with revisions in Everyman's University Library in 1973), 179-194.

Lycan, W. G. 2003, "Dretske's Ways of Introspecting" in *Privileged Access and First-Person Authority*, edited by B. Gertler (Burlington: Ashgate),

McCarthy, J. 1971-1987, "Formalization of Two Puzzles Involving Knowledge". manuscript available online at http://www-formal.stanford.edu/jmc/puzzles.html, first published in McCarthy [1990].

McCarthy, J. 1990, *Formalizing Common Sense: Papers by John McCarthy* (Norwood, NJ: Ablex).

Moody, T. C. 1994, "Conversations with Zombies", *Journal of Consciousness Studies*, 1(2), 196-200.

Nagel, T. 1974, "What Is It Like to Be a Bat?" *Philosophical Review*, 83(4), 435-450.

Rao, A., and Georgeff, M. 1991, "Modeling Rational Agents within a BDI-Architecture" in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, edited by J. Allen, R. Fikes, and E. Sandewall (San Mateo, CA: Morgan Kaufmann), 473-484.

Searle, J. R. 1992, *The Rediscovery of the Mind* (Cambridge, Mass; London: MIT Press).

Shimojo, S., and Ichikawa, S. 1989, "Intuitive Reasoning About Probability: Theoretical and Experimental Analyses of the "Problem of Three Prisoners"", *Cognition*, 32, 1- 24.

Symposium 1995, "Symposium on "Conversations with Zombies"", *Journal of Consciousness Studies*, 2(4).

Turing, A. M. 1950, "Computing Machinery and Intelligence", *Mind*, 59(236), 433-460.

Walton, D. N. 1991, "Critical Faults and Fallacies of Questioning", *Journal of Pragmatics*, 15, 337-366.

Werning, M. forthcoming, "Self-Awareness and Imagination" in *Mind and Action*, edited by J. Saagua.

Wooldridge, M. J. 2002, *An Introduction to Multiagent Systems* (Chichester: J. Wiley).