THINKING IMPOSSIBLE THINGS*

Sten Lindström

> "There is no use in trying," said Alice; "one can't believe impossible things." "I dare say you haven't had much practice," said the Queen. "When I was your age, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast".
> Lewis Carroll, *Through the Looking Glass*.
>
> As Ravi, aged 10, says: "It can be fun playing with ideas, like thinking impossible things and wondering if they are impossible."
> From Philosophy for children web page.

1. Believing the impossible

It is a rather common view among philosophers that one cannot, properly speaking, be said to believe, conceive, imagine, hope for, or seek what is impossible. Some philosophers, for instance George Berkeley and the early Wittgenstein, thought that logically contradictory propositions lack cognitive meaning (informational content) and cannot, therefore, be thought or believed. The view that what is thinkable – i.e., can be represented in thought – is also possible, was given a powerful expression in Wittgenstein's *Tractatus logico-philosophicus*. The fundamentals of Wittgenstein's view of the nature of thought are contained in the following sentences from *Tractatus*:

> 2.203. Das Bild enthält die Möglichkeit der Sachlage, die es darstellt.
> 2.221. Was das Bildt darstellt, ist sein Sinn.
> 2.222. In der Übereinstimmung oder Nichtübereinstimmung seines Sinnes mit der Wirklichkeit besteht seine Wahrheit oder Falschheit.

* I am grateful to Joseph Almog, Anders Berglund, Wlodek Rabinowicz, and the other contributors to this volume for helpful comments and stimulating discussions. I am likewise indebted to Krister Segerberg and Rysiek Sliwinski.

2.223. Um so erkennen, ob das Bild wahr oder falsch ist müssen wir es mit der Wirklichkeit vergleichen.

2.224. Aus der Bild allein ist nicht zu erkennen,ob es wahr oder falsch ist.

3. Das logische Bild der Tatsachen ist der Gedanke.

3.001. "Ein Sachverhalt ist denkbar" heisst: Wir können uns ein Bild von ihm machen.

3.02. Der Gedanke enthällt die Möglickeit der Sachlage, die er denkt. Was denkbar ist is auch möglich.

3.02. Wir können nichts Unlogisch denken, weil wir sonst unlogisch denken mussen.

Thus, thinking consists in the formation of pictures, i.e., isomorphic representations, of possible states of affairs (situations).[1] That a state of affairs is thinkable means that it can be represented (depicted) in thought. But only what is possible can be represented in thought. A picture of an impossible state of affairs would itself be an "illogical" configuration of thought constituents. But such a configuration would not make any sense. One might think that logically false (contradictory) sentences could do the job of representing impossible states of affairs. But such sentences, according to the early Wittgenstein, do not represent at all. A sentence that is logically true (a *tautology*) or logically false (a *contradiction*) is senseless, i.e., does not represent a possible state of affairs. If S is such a sentence, 'John believes that S', 'John imagines that S', etc. must also be senseless. The conclusion is that it is impossible to think something impossible.

Philosophers who do not go as far as Berkeley and Wittgenstein in denying that impossible propositions or states of affairs are thinkable, may still claim that it is impossible to *rationally believe* an impossible proposition. On a classical "Cartesian" view of belief, belief is a purely mental state of the agent holding true a proposition p that he "grasps" and is directly acquainted with. But if the agent is directly acquainted with an impossible proposition, then, presumably, he must know that it is impossible. But surely no rational agent can hold true a proposition that he knows is impossible. Hence, no rational agent can believe an impossible proposition.

A similar argument goes for conceivability and imaginability. On Stephen Yablo's analysis (Yablo, 1993), a proposition p is conceivable for an agent x if x can imagine a possible world which he takes to verify p. But if conceiving p is an inner mental state that involves the agent grasping or being acquainted with the proposition p, then it is hard to understand how a rational agent can

[1] I follow here the interpretation of Wittgenstein's views on representation and thought given by Eric Stenius. See Stenius (1964), Chapter VI, Section 13.

imagine a possible world that he takes to verify p, if p is not possible. Thus it seems that on the Cartesian view of propositional attitudes as inner mental states in which proposition are immediately apprehended by the mind, it is impossible for a rational agent to believe, imagine or conceive an impossible proposition.

Nowadays the Cartesian view of propositional attitudes is widely rejected. Even then, philosophers like Quine and Davidson, who are concerned with the hermeneutic process of ascribing beliefs and desires to agents in order to explain the way they talk and act, are reluctant to ascribe inconsistent, or otherwise impossible, beliefs to agents. This reluctance has to do with the role that the so-called *Principle of Charity* has in their theories of translation and interpretation. According to this principle, we interpret the talk and actions of another person in a way that maximizes the overall truth and coherence of the beliefs that we ascribe to the person. To ascribe inconsistent, or otherwise impossible beliefs, to an agent threatens to make him appear unintelligible. However, these philosophers do not impose an absolute ban on ascribing inconsistent, or impossible, beliefs to agents. It is instead a strong *desideratum* that agents not be described as having inconsistent beliefs. Davidson (2001) writes:

> ...unless one's beliefs are roughly consistent with each other, there is no identifying the contents of beliefs.
>
> This is not to say everyone is perfectly rational; anyone is capable of making a mistake in logic, or of entertaining beliefs that are inconsistent with each other. (Both Frege and Quine for example, wrote books based on inconsistent logics.) But there is a limit to how inconsistent a person can be and still be credited with clearly defined attitudes.

2. Alice's thesis

Ruth Barcan Marcus (1983) has suggested that a belief attribution is defeated once it is discovered that the proposition, or state of affairs. that is believed is impossible. According to her intuition, just as knowledge implies truth, belief implies possibility. It is commonplace that people *claim* to believe propositions that later turn out to be impossible. According to Barcan Marcus, the correct thing to say in such a situation is not: I once believed that A but I don't believe it any longer, since I have come to realize that it is impossible that A. What one should say is instead: It once appeared to me that I believed that A, but I did not, since it is impossible that A.

Barcan Marcus (1983) writes:

> "...there is a still stronger view, which I would like to defend: Alice's view, that we cannot believe an impossibility. ... The position I propose is that, just as knowing that p relates an agent to an actual state (or states) of affairs, otherwise the knowledge claim [the claim that one knows] is mistaken, so there is an important sense of 'believes' such that believing p relates an agent to a possible state (or states) of affairs, otherwise the belief claim [the claim that one believes] is mistaken despite the apparent evidence to the contrary.

Thus, Barcan Marcus defends what we might call *Alice's thesis*:

> *Necessarily, for any proposition p and any subject x, if x believes p, then p is possible.*

In symbolic form, this becomes:

Alice's thesis: $\mathbf{L}\forall p\forall x(\mathbf{B}_x(p)\rightarrow \mathbf{M}p)$.

Or, equivalently:

$$\neg\mathbf{M}\exists p\exists x[\mathbf{B}_x(p) \wedge \neg\mathbf{M}p].$$

That is, it is not possible that there should be a proposition p and an agent x such that x believes p and p is impossible.

Alice's thesis, that it is impossible to hold impossible beliefs, seems to come into conflict with our ordinary practices of attributing beliefs. Consider a mathematician who believes all the axioms $A_1,..., A_n$ of a certain mathematical theory together with the negation of one of its theorems B. We have the following situation:

(1) $A_1 \wedge A_2 \wedge...\wedge A_n$ logically implies B.

(2) $\mathbf{B}_m(A_1 \wedge A_2 \wedge...\wedge A_n \wedge \neg B)$.

Since logical relations hold necessarily, it follows from (1) that:

(3) $\mathbf{L}(A_1 \wedge A_2 \wedge...\wedge A_n \rightarrow B)$.

That is,

(4) $\neg\mathbf{M}(A_1 \wedge A_2 \wedge...\wedge A_n \wedge \neg B)$.

Thus, our mathematician is apparently believing an impossible proposition.

Consider another mathematical example. Some mathematicians believe that CH (*the continuum hypothesis*) is true and others believe that it is false. But if CH is true, then it is necessarily true; and if it is false, then it is necessarily false. Regardless of whether CH is true or false, the conclusion seems to be that there are mathematicians who believe impossible propositions.

Examples of apparent beliefs in impossible propositions outside of mathematics are also easy to come by. Consider, for example, Kripke's (1999) story of

the Frenchman Pierre who without realizing it has two different names 'London' and 'Londres' for the same city, London. After having arrived in London, Pierre may *assent* to 'Londres is beautiful and London is not beautiful' without being in any way irrational. It seems reasonably to infer from this that Pierre believes that Londres is beautiful and London is not beautiful. That is,

(1) $\mathbf{B}_{\text{Pierre}}$(Londres is beautiful $\wedge$ $\neg$London is beautiful).

But since 'Londres' and 'London' are rigid designators for the same city, the following must also be true:

(2) $\neg\mathbf{M}$(Londres is beautiful $\wedge$ $\neg$London is beautiful).

So Pierre appears to believe an impossible proposition.

Finally, let us consider an example from philosophy of apparent belief in impossible propositions. Many philosophers seem to have believed that knowledge is justified true belief, so for some x it seems to be the case that:

(1) $\mathbf{B}_{x}$(knowledge = justified true belief).

Suppose now that Edmund Gettier (1963) is correct and:

(2) knowledge $\neq$ justified true belief.

Then, presumably:

(3) $\neg\mathbf{M}$(knowledge = justified true belief).

Hence, at least some philosophers seem to have believed impossible propositions.

3. A transcendental argument

I am here going to present an argument, adopted with some modifications from Roy Sorensen (1996), for the claim that it is possible to believe impossible propositions.

Consider:

Alice's thesis: $\mathbf{L}\forall p\forall x(\mathbf{B}_{x}(p) \rightarrow \mathbf{M}p)$.

As we have seen, there are apparent counterexamples to the thesis. The proponents of the thesis, for instance Barcan Marcus (1983), claim that these counterexamples are merely apparent. However, the proponents of the thesis do not deny that there are philosophers who disbelieve their thesis. On the contrary, they are ready to argue with those philosophers and try to convince them that

they are wrong. Hence, the following proposition seems to be an uncontested fact:

(1) $\exists x \mathbf{B}_x(\neg$Alice's thesis),

i.e., there are people who disbelieve Alice's thesis.

We consider it to be a condition of adequacy on any theory of belief that it be compatible with the assumption (1).

From (1) we get by existential instantiation:

(2) $\mathbf{B}_a(\neg$Alice's thesis)

Suppose now, for a reductio, that Alice's principle is true, i.e.,

(3) $\mathbf{L}\forall p \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{M}p)$.

(2) and (3) yield:

(4) $\mathbf{M}\neg$Alice's thesis.

In other words:

(5) $\mathbf{M}\neg\mathbf{L}\forall p \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{M}p)$.

But since the logic of metaphysical necessity is at least as strong as S4, we can infer from (5):

(6) $\neg\mathbf{L}\forall p \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{M}p)$.

That is,

(7) $\neg$Alice's thesis.

(3) and (7) contradict each other. Hence, by Reductio ad Absurdum,

(8) $\neg$Alice's thesis.

Thus, given the fact that there are people who disbelieve Alice's thesis, we can conclude that Alice's thesis is indeed false. Hence, we have what one might think of as a *transcendental argument* for the falsity of Alice's thesis.

However, we can go on. Let us conditionalize on the assumption (1). We then get a logical proof of:

(9) $\exists x \mathbf{B}_x(\neg$Alice's thesis) $\rightarrow \neg$Alice's thesis.

Applying necessitation to (9), we get:

(10) $\mathbf{L}(\exists x \mathbf{B}_x(\neg$Alice's thesis) $\rightarrow \neg$Alice's thesis).

From this we can infer:

(11) $\mathbf{M}\exists x \mathbf{B}_x(\neg$Alice's thesis) $\rightarrow \mathbf{M}\neg$Alice's thesis.

But,

(12) $\mathbf{M}\neg$Alice's thesis $\rightarrow \neg$Alice's thesis.

From (11) and (12) we conclude:

(13) $\mathbf{M}\exists x\mathbf{B}_x(\neg$Alice's thesis$) \rightarrow \neg$Alice's thesis.

That is, if it is possible to disbelieve Alice's thesis, then the thesis is false.

Now, the proponents of Alice's thesis seem to accept the following two claims:

(14) Alice's thesis.

(15) $\mathbf{M}\exists x\mathbf{B}_x(\neg$Alice's thesis$)$.

But if they do, they are inconsistent (by 13).

4. Hume's conceivability-thesis

Let us now turn to the so-called Hume's thesis that conceivability implies possibility:[2]

Hume's thesis:
Necessarily, for any proposition p, if p is conceivable, then p is possible.

Writing $\mathbf{C}_x(p)$ for 'The proposition p is conceivable for x, we can rephrase this as:

Hume's thesis: $\mathbf{L}\forall p\forall x[\mathbf{C}_x(p) \rightarrow \mathbf{M}p]$.

There is also the following stronger version of Hume's thesis:

Hume's strong thesis: $\mathbf{L}\forall p\forall x[\mathbf{C}_x(p) \leftrightarrow \mathbf{M}p]$.

Suppose now that the proponents of Hume's thesis accept that there could exist an agent for whom it is conceivable that Hume's thesis is false. Suppose that **a** is such a proponent. Then, **a** accepts:

(1) Hume's thesis.

(2) $\mathbf{M}\exists x\mathbf{C}_x(\neg$Hume's thesis$)$.

From (1) we can infer:

(3) $\mathbf{L}\forall p[\mathbf{M}\exists x\mathbf{C}_x(p) \rightarrow \mathbf{M}p]$.

---

[2] For a thorough discussion of this thesis, various interpretations of it, and its philosophical plausibility see Anders Berglund's contribution to this volume.

From (2) and (3) we infer:

(4) **M**¬Hume's thesis.

But,

(5) Hume's thesis → **L**Hume's thesis.

From (1) and (5), we get:

(6) **L**Hume's thesis.

But (4) and (6) yield a contradiction.

Thus, if one accepts Hume's thesis one cannot, on pain of contradiction, accept the thesis that there could be an agent for whom it is conceivable that the thesis is false.

We can of course also formulate the result as:

(7) Hume's thesis → ¬**M**∃x$\mathbf{C}_x$(¬Hume's thesis).

That is, if it is conceivable that Hume's thesis is false, then it is false.

The proponents of Hume's thesis might respond that this argument by pointing out that it is a simple consequence of Hume's thesis that:

(8) **L**∀p[**L**p → ¬**M**∃x$\mathbf{C}_x$(¬p)].

(8) is just a easy consequence of (3) above. From (8) we get (7) since Hume's thesis is necessarily true. So the proponent of Hume's thesis might want to bite the bullet and accept (7) and (8).

However, (8) is extremely unintuitive. Consider for instance, Andrew Wiles's proof of Fermat's last theorem. Assuming that the proof is valid, it is necessarily valid. If P is a valid proof of a mathematical statement S, it is necessarily true that P is a valid proof of S. However, it seems quite obvious, due to the extremely complicated nature of the proof, that even a highly skilled expert in the field (Wiles himself, for instance) can still imagine that there is some mistake in the proof. So even if the proof is necessarily valid, it is conceivable that it is not.[3] The conclusion is that Hume's thesis, in its general form, is extremely unintuitive.

[3] This kind of argument appears in Putnam (1990).

5. The problem of logical omniscience

The attempts to provide propositional attitude reports with a model-theoretic semantics of the possible worlds variety goes back to Jaakko Hintikka's seminal work *Knowledge and Belief* (1962). I shall refer to this approach as the c*lassical possible worlds approach* to belief and other propositional attitudes. The intuitive idea behind this approach is to think of the agent's (current) state of belief as dividing the set of possible scenarios ("possible situations", "possible worlds") into two classes consisting of the scenarios that are compatible with the agent's state of belief and those that are not. According to Hintikka to attribute a belief to a person x is to invoke the idea of a set of "doxastically" possible scenarios (with respect to the person x). These scenarios, the person's *doxastic alternatives,* are precisely the scenarios (situations, worlds) that are compatible with everything that the person believes (in the "actual" scenario).[4] Although they all agree with respect to what the person believes, they still differ in ways that make them incompatible with each other. The analogy with necessity leads to the following truth clause for belief:

> 'x *believes that* A' is true in the scenario u if and only if, in every possible scenario v compatible with what x believes in u, A is true in v.

Or more briefly put:

> 'x *believes that* A' is true in u if and only if, for every doxastic x-alternative v to u, A is true in v.

Or shorter still:

> $u \vDash \mathbf{B}_x(A)$ iff for every v, if $u\ B_x\ v$, then $v \vDash A$.

To be more precise, let us define a *classical doxastic model* as a structure M = <S, W, I, B, @> consisting of: (i) a set S, the elements of which are called *possible scenarios* (or *points*); (ii) W is a subset of S, the elements of which are

---

[4] Although, Hintikka speaks of his points of evaluation as "possible worlds", it may be misleading to refer to them by that term. Nowadays, "possible worlds" are usually thought of as metaphysically possible worlds in the sense of Kripke. This is definitely not the way one should think of Hintikka's possible worlds. It is better to think of Hintikka's worlds as models (in the sense of model theory), state descriptions, or maximal consistent sets of sentences. Actually, he works with so-called *model sets*, i.e., downwards saturated sets of sentences (see Hintikka, 1962, 1969). That is his worlds are required to be (negation-)complete and formally consistent, but they do not need to respect the metaphysical constraints of genuine possible worlds. Coreferring individual constants ("proper names") may even designate distinct individuals in some of Hintikka's worlds. I prefer to refer to them as *scenarios* rather than *worlds*.

called *possible worlds*; (iii) a set I, the elements of which are called *agents*; (iv) a function B assigning to each member x of I a *doxastic alternativeness relation* $B_x \subseteq S \times S$; and finally (v) a designated point called the *actual world*. We assume that $@ \in W$, so that @ is indeed a world. By an *assignment* g, we understand a function which assigns elements of I to the individual variables x_1, x_2,... and subsets of S to the propositional variables p_1, p_2, ...[5] Given a model M = <S, W, I, B, @>, an assignment g and a point u, we can define for any formula A of L what it means for A to be *satisfied* in M by u and g (in symbols: M, u, g $\models$ A). Here are some of the defining clauses:

(i) M, u, g $\models$ p iff $u \in g(p)$.

(ii) M, u, g $\models$ (p = q) iff g(p) = g(q).

(iii) M, u, g $\models$ **L**A iff for all $v \in S$, M, v, g $\models$ A.

Here, **L** is the operator of *logical necessity*. Remember that according to the intended interpretation there are metaphysically possible as well as metaphysically impossible (although) scenarios in S.

(iv) M, u, g $\models$ **L**A iff $u \in W$ and for all $v \in W$, M, v, g $\models$ A

L is the operator of metaphysical necessity. **L**A is true at a point u iff u is a possible world and A is true at all possible worlds. Thus, **L**A is by stipulation false at every scenario which is not a possible world. We get the following clause for **M**:

(v) M, u, g $\models$ **M**A iff $u \notin W$ or there is a $v \in W$, M, v, g $\models$ A.

(vi) M, u, g $\models$ $\mathbf{B}_x$(A) iff for all $v \in S$, if $u\ B_{g(x)}\ v$, then M, u, g $\models$ A.

(vii) M, u, g $\models$ $\forall x A(x)$ iff for all $a \in I$, M, u, g(a/x) $\models$ A(x), where g(a/x) is the assignment which is exactly like g except for assigning a to x,

(viii) M, u, g $\models$ $\forall p A(p)$ iff or all $X \subseteq S$, M, u, g(X/p) $\models$ A(p), where g(X/p) is the assignment which is exactly like g except for assigning X to p.

---

[5] Although I have not given a formal definition of the object language L, it should be fairly clear by now what kind of language that I have in mind. It is a language with propositional variables, Boolean sentential connectives, propositional quantifiers, individual variables x_1, x_2,... , propositional variables p_1, p_2,... , individual and propositional quantifiers, operators $\mathbf{B}_x$, $\mathbf{C}_x$,... for the various propositional attitudes, operators $\Box$ and $\Diamond$ for metaphysical necessity and possibility, and operators **L** and **M** for logical necessity and possibility. We shall also let the language L contain the symbol = both as a sign for identity between agents as a sign for identity between propositions.

We say that A is *true* at the point u in M (in symbols, M, u $\models$ A) iff it holds for every assignment g that M, u, g $\models$ A. A is *true* in the model M (in symbols, M $\models$ A) iff M, @ $\models$ A, i.e. iff A is true at the actual scenario @ in M. A is *logically valid* (in symbols, $\models$ A) iff A is true in every model M.

Since the actual scenario is a possible world (i.e., belongs to W), we have the following clauses for **L** and **M** at @:

(ix) M, @, g $\models$ **L**A iff for all v $\in$ W, M, v, g $\models$ A.

(x) M, @, g $\models$ **M**A iff there is a v $\in$ W, M, v, g $\models$ A.

Hence, the logic for **L** and **M** is the usual S5.

The classical semantics for doxastic (and epistemic) logic gives rise to a number of properties of the resulting logic that go under the name properties of *logical omniscience.* We consider these properties for belief, but analogous properties hold on the classical approach also for knowledge and other propositional attitudes:[6]

LO 1: $\models \forall p \forall x[B_x(p \rightarrow q) \rightarrow (\mathbf{B}_x(p) \rightarrow \mathbf{B}_x(q))]$,

i.e., if an agent believes that p $\rightarrow$ q and he believes p, then he believes q.

LO 2: $\models \forall p \forall x[\mathbf{L}p \rightarrow \mathbf{B}_x(p)]$,

i.e., every agent believes every logically necessary proposition.

LO 3: $\models \forall p[\mathbf{L}(p \rightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{B}_x(q))]$,

i.e., if p logically necessitates q, then every agent that believes p also believes q. In other words, an agent's beliefs are closed under logical necessitation.

LO 4: $\models \forall p \forall q[\mathbf{L}(p \leftrightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \leftrightarrow \mathbf{B}_x(q))]$,

i.e., if p and q are logically equivalent propositions, then no agent can believe p without believing q and vice versa.

LO 5: $\models \forall p \forall x[\mathbf{B}_x(p \wedge q) \leftrightarrow (\mathbf{B}_x(p) \wedge \mathbf{B}_x(q))]$,

i.e., an agent believes the proposition p $\wedge$ q iff he believes both p and q.

LO 6: $\models \forall p \forall q \forall x[(B_x(p) \wedge \neg\mathbf{M}p) \rightarrow \mathbf{B}_x(q)]$.

i.e., if an agent believes an logically impossible proposition, then he believes everything.

As an illustration, let us verify that the property LO 6 is true in every model. Suppose that:

---

[6] Compare Meyer (2001).

(1) $M, @, g \models (B_x(p) \wedge \neg \mathbf{M}p)$.

But then g(p) must be the empty set (the logically impossible proposition) and g(p) must be true in all doxastic g(x)-alternatives to @. But the logically impossible proposition ∅ is not true in any scenario, so @ cannot have any doxastic g(x)-alternatives. But then g(q) will be vacuously true at all doxastic g(x)-alternatives to @. Hence,

(2) $M, @, g \models B_x(q)$

Thus, LO 6 must be true in every model M.

Obviously, no agent can believe every proposition, so if we impose the requirement that $\forall a \in I, \forall u \in S, \exists v \in S, u\ B_a\ v$ (Seriality), then

LO 7: $\models \forall x \forall p(B_x(p) \rightarrow \mathbf{M}p)$.

That is, logically impossible proposition cannot be believed.

If we assume Seriality, then we also have::

LO 8: $\models \forall p \forall x \neg(\mathbf{B}_x(p) \wedge \mathbf{B}_x(\neg p))$.

i.e., no agent can believe both p and ¬p.

We do not have the properties for metaphysical necessity and possibility that correspond to LO 3, LO 4 and LO 6 and LO 7. That is, none of the following sentences are valid:

$\forall p \forall x[\mathbf{L}p \rightarrow \mathbf{B}_x(p)]$

$\forall p[\mathbf{L}(p \rightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{B}_x(q))]$

$\forall p \forall q[\mathbf{L}(p \leftrightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \leftrightarrow \mathbf{B}_x(q))]$

$\forall p \forall q \forall x[(B_x(p) \wedge \neg \mathbf{M}p) \rightarrow \mathbf{B}_x(q)]$

$\forall x \forall p(\mathbf{B}_x(p) \rightarrow \mathbf{M}p)$.

So even if it is true (in the actual world) that **L**(Hesperus = Phosphorus), it does not follow that $\mathbf{B}_{\text{Jones}}$(Hesperus = Phosphorus). In some of Jones's doxastic scenarios, 'Hesperus' and 'Phosphorus' may very well pick out different entities. This can be so, even if we assume that 'Hesperus' and 'Phosphorus' designate Venus in all possible worlds. It may even be the case that $\mathbf{B}_{\text{Jones}}$(Hesperus ≠ Phosphorus). Thus, it is important to distinguish between (metaphysically) possible worlds and epistemic scenarios.

Using this distinction we may be able to explain how someone might come to believe the negation of a true mathematical theorem. Let us assume for instance that CH (the continuum hypothesis) is in fact false in the real set-theoretic universe (the full cumulative hierarchy of sets) **V**. Then CH is false in every

possible world. All the possible worlds have the same universe of pure sets, which is identical to **V**. In the agent's belief scenarios, however, the universe of sets could be some other well-behaved model of Zermelo-Fraenkel set theory, e.g. the hierarchy **L** of constructible sets, where CH is true. In this way, we may be able to explain how a fully rational agent can come to believe a necessarily false mathematical statement. This is just a sketch of a strategy to alleviate some of the problems of logical omniscience. The suggestion obviously has to be worked out in detail.

However, we still have the problem of logical omniscience in the limited sense of LO 1 – LO 8. This is bad enough. Presumably all the theorems of predicate logic will hold in all epistemic scenarios, but the agent cannot – in any realistic sense – be expected to believe them all. It should be possible for an agent to believe A while not believing B, although B is a first-order consequence of A.

Let us analyze the problem of logical omniscience. The problem arises as soon as we assume:

LO 4: $\models \forall p \forall q[\mathbf{L}(p \leftrightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \leftrightarrow \mathbf{B}_x(q))]$

(*Distribution*): $\models \forall p \forall x[(\mathbf{B}_x(p \wedge q) \rightarrow (\mathbf{B}_x(p) \wedge \mathbf{B}_x(q))]$.

together with the assumption no agent can believe every proposition, that is:

(*Nontriviality*) $\models \neg \mathbf{M} \exists x \forall p \mathbf{B}_x(p)$.

From these three principles, we can infer both:

LO 3: $\models \forall p[\mathbf{L}(p \rightarrow q) \rightarrow \forall x(\mathbf{B}_x(p) \rightarrow \mathbf{B}_x(q))]$

and

LO 7: $\models \forall x \forall p(\mathbf{B}_x(p) \rightarrow \mathbf{M}p)$.[7]

Since the latter two principles are clearly unacceptable, we need to give up at least one of LO 4, Distribution or Nontriviality. If there are *any* reasonable principles of doxastic logic, Distribution seems to be one of them. The same goes for Nontriviality. Thus, it seems that LO 4 has be given up. That is, we

---

[7] Informal proof of LO 3: Suppose $\mathbf{L}(p \rightarrow q)$ and that $\mathbf{B}_x(p)$.. Then, it is true in every scenario that $p \rightarrow q$. Hence, it is true in every scenario that $p \wedge q \leftrightarrow p$. Thus it is true that $\mathbf{L}(p \leftrightarrow p \wedge q)$. So by LO 4 , $\mathbf{B}_x(p) \leftrightarrow \mathbf{B}_x(p \wedge q)$. Thus $\mathbf{B}_x(p \wedge q)$. Finally, we get $\mathbf{B}_x(q)$ using Distribution.

Informal proof of LO 7: Suppose for reductio that $\mathbf{B}_x(p)$ and $\neg \mathbf{M}p$. Then there cannot be any doxastic x-alternatives to the actual world. But the agent will believe any proposition (in the actual world), contrary to Nontriviality.

should give up the principle that logically equivalent propositions are interchangeable in belief contexts.

One suggestion is to add some kind of "fine grained" propositions to the modeling. A model would then be a structure M = <S, W, I, PROP, «», B, @>, where the new ingredients are: (i) a set PROP of "structured propositions" and (ii) a function «» assigning to each proposition **p** in Prop an appropriate intension «**p**» ⊆ S. An assignment now assigns individuals to the individual variables and members of PROP to the propositional variables. The operators **L** and **L** operate on intensions as before, but the belief operators $\mathbf{B}_x$, as well as other propositional operators, are "ultraintensional" operators on structured propositions. The principle LO 4 is now replaced by:

$$\models \forall p \forall q[(p = q) \rightarrow \forall x(\mathbf{B}_x(p) \leftrightarrow \mathbf{B}_x(q))],$$

i.e., if the propositions p and q are identical, then $\mathbf{B}_x(p)$ holds iff $\mathbf{B}_x(q)$. This is of course a consequence of belief being a propositional operator, i.e., an operator from PROP to PROP.

However, we do not just want to introduce fine-grained propositions. We would also like to preserve something of the classical possible worlds analysis of belief. This can be done by following the lead of Fagin and Halpern's (1988) so-called *sieve-semantics*.[8] For any sentence A, let «A» be the proposition that A expresses in the given model. We now add still another ingredient to the model: For each scenario x and each agent **a**, there is a set **A**(**a**, u) (the "sieve") of all the propositions that the agent **a** is *aware* of in u. We can now give the following new semantic clause for the belief operator:

> M, u, g ⊨ $\mathbf{B}_x$(A) iff (i) «A» ∈ **A**(g(x), u); and (ii) for all v ∈ S, if u $B_{g(x)}$ v, then M, u, g ⊨ A.

The intuitive meaning of this is: $\mathbf{B}_x$(A) is true in the scenario u iff (i) x is in u aware of the proposition expressed by A; and (ii) A is true in all doxastic x-alternatives to u. Thus **A**(**a**, u) functions like a sieve: it eliminates from the agent's beliefs propositions that he is not aware of.

In the sieve semantics, the principles LO 1 – LO 6 are no longer valid. However, in order to preserve some additional logical principles, one may want to assume, for example, that the agents are aware of (all substitution instances of) tautologies of sentential logic, but not necessarily of all the theorems of predicate logic. We would then have:

> If A is valid in sentential logic, then ⊨ $\forall x\mathbf{B}_x$(A)

---

[8] See also Meyer (2001) and Fagin, Halpern, Moses, and Vardi (1975), Chap. 9.

If $(A \leftrightarrow B)$ is a tautology, then $\models \forall x(\mathbf{B}_x(A) \leftrightarrow \mathbf{B}_x(B))$.

But we would not have the analogous principles for valid formulas of predicate logic.

One could of course idealize further by assuming that the agents are aware of all formulas of some decidable fragment of first-order predicate logic.

We may also want to assume:

If «$A \wedge B$» $\in$ $\mathbf{A}(\mathbf{a}, u)$, then «$A$» $\in$ $\mathbf{A}(\mathbf{a}, u)$ and «B» $\in$ $\mathbf{A}(\mathbf{a}, u)$,

i.e., if an agent is aware of a conjunctive proposition, then he is also aware of each of the conjuncts of the proposition. This assumption makes the Distribution principle valid.

The sieve semantics, however, still keeps the principles:

LO 7: $\models \forall x \forall p(B_x(p) \rightarrow \mathbf{M}p)$

LO 8: $\models \forall p \forall x \neg(\mathbf{B}_x(p) \wedge \mathbf{B}_x(\neg p))$.

given that the alternativeness relations are serial. LO 7 forbids belief in contradictions. LO 8, on the other hand, forbids contradictory beliefs. We may want to preserve LO 7 while giving up LO 8.

One intuitive reason while LO 8 might fail is that the agent may believe p in one frame of mind, while believing $\neg p$ in another frame of mind. We may imagine that the agent has received the information p from one source and the information $\neg p$ from another, but has not yet put the two pieces of information together. If we interpret $\mathbf{B}_x(A)$ as meaning that there is some frame of mind f such that x believes A in x, then $\mathbf{B}_x(p) \wedge \mathbf{B}_x(\neg p)$ may very well be true. This approach is pursued by Fagin and Halpern (1988) in their so called *cluster semantics* (also called semantics for *local reasoning*). The general idea is to equip the agent with a family of alternativeness relations, one for each frame of mind. Each frame of mind may be consistent, so the agent will never believe a contradiction. However contradictory beliefs are allowed.

Let us now summarize. The classical possible worlds analysis of belief is often criticized for its apparent commitment to *logical omniscience*, i.e., the assumptions that (i) everyone believes all logical truths, (ii) everyone believes all the logical consequences of what he believes; and (iii) no one can ever believe a logically false proposition. The commitment to logical omniscience becomes even worse, if the possible worlds of doxastic logic are identified with metaphysically possible worlds in the sense of Kripke. Then the possible worlds analysis seems to commit us not only to logical omniscience but even worse to *modal omniscience:* (i) everyone believes every (metaphysically) necessary

statement; (ii) everyone's beliefs is closed under necessary implication; (iii) nobody can ever believe anything impossible. In order to avert the threat of modal omniscience, we suggested that one should make a distinction between (epistemically) possible scenarios and metaphysically possible worlds. The collection of metaphysically possible worlds is a proper subset of the set of all epistemically possible scenarios. We showed how this distinction could be used to explain the failure of agents to believe all mathematical truths and their failure to believe metaphysical necessities that are not logically necessary It was also explained how agent's could come to believe aposteriori impossibilities like Hesperus ≠ Phosphorus. An agent might simply not have enough information to know (or believe) that 'Hesperus' and 'Phosphorus' pick out the same object. But then the two names will pick out different objects in at least some of the scenarios that are compatible with what the agent believes. It may even be that the two names pick out different objects in all scenarios compatible with what the agent believes.

Once the threat of modal omniscience has been avoided, it remained to explain why ordinary agent's may not even be logically omniscient in the narrow sense. As long as the objects of propositional attitudes are taken to be truth-supporting scenarios or worlds, the assumption of logical omniscience can hardly be avoided. We therefore suggested that a more fine-grained notion of "structured" propositions was needed. No analysis was given of the structure of propositions, but it was suggested that propositions must be taken to have a syntactic structure of a sentence-like character. The analysis of belief as relation to propositions need not mean that the agent has direct acquaintance with the objects of his own beliefs. However, in order to preserve some of the advantages of the possible world analysis we followed Fagin and Halpern (1988) in analyzing (explicit) belief into two components that we may speak of as: awareness and doxastic commitment. An agent (explicitly) believes a proposition iff he is aware of the proposition and he is doxastically committed to its truth. The second component, doxastic commitment, was analyzed along the lines of Hintikka-type possible worlds semantics. Finally we discussed ways in which it is possible to have contradictory beliefs without believing contradictions. An agent may have contradictory beliefs, if his thinking is compartmentalized into different frames of mind. He may believe p in one frame of mind and ¬p in another, without realizing that he has contradictory beliefs.

6. Conceiving of one thing as two: a puzzle about de re conceivability

In this section I am going to consider the question whether it is possible to believe, imagine, or conceive what is impossible from a slightly different perspective. For one thing, I will consider conceivability (imaginability) rather than belief. The discussion is carried out with a view to conceivability arguments in the philosophy of mind. Moreover, the discussion will concern attitudes *de re* rather than *de dicto*. The question in focus will be: Is it possible to conceive of one thing as two. For instance, is it possible to conceive of mind and body as distinct entities, even if they are inseparable.

6.1. Conceivability arguments

Let a conceivability argument be any argument from the conceivability of a thing or state of affairs to its possibility or actuality. Conceivability arguments are ubiquitous in the philosophy of mind. In this note, I am going to focus on one particular kind of conceivability arguments, namely, arguments from the conceivability of distinctness to real distinctness. These are arguments from the separability in thought of two phenomena to their distinctness in reality. Their guiding principle, or what we might call the *separability-by-conceivability principle*, was clearly stated by Descartes:

> "...the fact that I can clearly and distinctly understand one thing without another is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God." (Descartes, *Meditations on First Philosophy, Meditation VI*)

Descartes argued: If we can clearly and distinctly conceive of a situation in which one thing A exists but another thing B doesn't, then it is possible for A to exist without B existing. But then A and B must be two distinct entities. Descartes famously applied this kind of reasoning to the Mind and the Body: Since we can (clearly and distinctly) conceive of the Mind without the Body (or existing together with another Body), it is indeed possible for the former to exist without the latter. Hence, the two must be distinct.

An argument of a similar kind separating the mental from the physical has more recently been put forward by Saul Kripke in *Naming and Necessity*. In rough outline, Kripke's argument goes as follows: Suppose that a given phenomenal state-type X (pain, for instance) turns out to be perfectly correlated with a certain physical state-type Y (usually thought of as some kind of brain state that philosophers are fond of misdescribing as "C-fiber stimulation"). Some philosophers may argue that the best explanation of such a correlation is that X and Y are one and the same state. It seems however, that we can

conceive of (or imagine) a situation in which the phenomenal state X occurs but the corresponding physical state Y does not (or vice versa). And conceivability, it is thought, is at least prima facie evidence for possibility. Hence it appears possible for X to exist without Y. If X can exist without Y, then X and Y must be distinct.

In the course of his argument, Kripke has to meet two possible objections.

When we conceive of X and Y, we must do so by means of some (linguistic or mental) representations, say R_X and R_Y. But if R_X and R_Y could fail to pick out their actual referents, then the following inference would not go through

(1) "$R_X \neq R_Y$" is possibly true. Hence, "$R_X \neq R_Y$" is (actually) true.

To make this kind inference, would then be to commit the same fallacy as in:

(2) "The Morning Star ≠ The Evening Star" is possibly true. Hence, "The Morning Star ≠ The Evening Star" is (actually) true.

This first objection is countered by Kripke by means of his claim that terms like "pain" and "C-fiber stimulation" are *rigid designators*. If R_X and R_Y are rigid designators that refer to X and Y necessarily ("in all possible worlds"), then the inference (1) is valid.[9]

The second objection against Kripke's conceivability argument is directed against the step:

(3) "$R_X \neq R_Y$" is conceivably true. Hence, "$R_X \neq R_Y$" is possibly true.

Or to take Kripke's own example:

(4) "Pain ≠ C-fiber stimulation" is conceivably true. Hence, "Pain ≠ C-fiber stimulation" is possibly true.

Kripke admits that inferences like (4) are defeasible. The (apparent) conceivability of a statement S is, at best, defeasible evidence for its truth.

A critic might argue that even if "$R_X \neq R_Y$" is true, and hence necessarily true, it might *appear* to us that the statement could have been false. The appearance of possibility could be due to a kind of *cognitive illusion*: due to our lack of knowledge of the true nature of pain. It may seem possible that pain be distinct from C-fiber stimulation, although in reality it is not. But says the critic: If pain is in fact C-fiber stimulation, then we could not, if we were fully

---

[9] Suppose R_X and R_Y are rigid designators referring to X and Y respectively. Suppose further that "$R_X \neq R_Y$" is possibly true. Then, there is a possible world in which it is true that X ≠ Y. That is, it is true of Y that it could have been distinct from X. However, it is not true of X that it could have been distinct from X. Hence Y has a property that X does not have. Hence, X ≠ Y, by the indiscernibility of identicals (Leibinz's Law).

informed of the nature of pain, conceive of it without C-fiber stimulation. An analog: As long as we don't know that water = H_2O, it is (apparently) conceivable that there should be water present but no H_2O. But the appearance of possibility disappears as soon as we learn that water = H_2O. When we are fully informed, we understand that the apparent possibility of water being different from H_2O had its source in the real possibility of the appearance of water ("the watery stuff") being something else than H_2O. So the apparent conceivability of "Pain ≠ C-fiber stimulation" can, according to the critic be explained in an analogous way as being due to our lack of knowledge of the nature of pain. According to the critic, what one really conceives when one seems to conceive of a situation where "Pain ≠ C-fiber stimulation" is true, is just an epistemic situation that is subjectively indiscernible from the real situation (where Pain is C-fiber stimulation) and where one *feels* pain without there being any stimulation of C-fibers. What this shows according to the critic is only that:

(5) The appearance of pain ≠ C-fiber stimulation,

not that:

(6) Pain ≠ C-fiber stimulation.

Against this argument Kripke objects that nothing could be pain if it did not feel like pain. Thus,

(7) The appearance of pain = pain.

For phenomenal states there is no distinction between what the state really is and how it appears to the person being in the state. Hence, if the appearance of X is possible without the state Y, then X itself is possible without Y.

According to Kripke, the critic's attempt to explain away our intuition that pain can occur without its physical correlate has failed. The critic has failed in his attempt to defeat the argument. Having, rebutted the objections to his conceivability argument, Kripke (tentatively) concludes that phenomenal states cannot be identical (or identified with) physical states.

6.2. "New Wave" materialism

Kripke's conceivability argument can be presented in the following simplified form:

(1) It is (apparently) conceivable that pain exists and C-fiber stimulation does not exist.

(2) If a state of affair is apparently conceivable and we cannot "explain away" its apparent conceivability, then it is reasonable to believe that the state of affair is really possible. (Conceivability is defeasible evidence for possibility).

(3) The usual way of explaining away apparent conceivability does not work in the case of pain and C-fiber stimulation.

(4) Hence, we seem to have good reasons to believe that it is possible that pain exists without C-fiber stimulation.

(5) If it is possible that pain exists without C-fiber stimulation, the pain is distinct from C-fiber stimulation.

(6) We seem to have good reasons to believe that pain is distinct from C-fiber stimulation.

In a recent paper, Horgan and Tienson (2001) have coined the term "New Wave Materialism" for a different way of meeting Kripke's challenge to materialism. The new wave materialists accept the view that phenomenal states are identical to (broadly) physical states. Furthermore they accept the view that all identities are necessarily true. They reject, however, Kripke's conceivability-possibility principle and his requirement that the apparent conceivability be "explained away". In the case of phenomenal and physical properties the (apparent) conceivability of separability is no guide to possibility; not even a defeasible guide. Instead it is a natural consequence of the different roles of phenomenal and physical concepts that the phenomenal properties should appear to be distinct from physical properties.

Phenomenal concepts are according to the new wave position (i) applied on the basis of introspection, (ii) refer directly and rigidly to phenomenal properties, and (iii) present the properties directly as they are in themselves. When one conceives of "Pain ≠ C-fiber stimulation" being true, one imagines the property of pain via its associated phenomenal concept and conceives of C-fiber stimulation via a physical concept. Due to the radical difference between the two kinds of concepts there is no way of knowing *a priori* whether they pick out the same property or not. Moreover the huge epistemic difference between introspecting, or imagining pain, and observing, or conceiving C-fiber stimulation gives rise to the appearance of distinctness between the properties. This appearance may be there regardless of whether we believe that the properties are the same or not. In this respect the appearance of distinctness is similar to a perceptual illusion like the Müller-Lyer illusion.

6.3. Conceivability de dicto and de re

It is common to distinguish between *de dicto* and *de re* modalities and attitudes.[10] In particular, this distinction can be applied to conceivability and imaginability. For instance,

(1) John imagines that: Water ≠ H_2O

is a *de dicto* report, while

(2) John imagines of the substances Water and H_2O that they are distinct,

is de re. The latter report can also be formulated as:

(3) The substances Water and H_2O are imagined by John to be distinct.

The common way of representing (1) and (2) in logical notation is as:

(1') Imagines(John, Water ≠ H_2O); and

(2') $\exists x \exists y(x = \text{Water} \wedge y = H_2O \wedge \text{Imagines}(\text{John}, x \neq y))$,

respectively.

Applying the *de dicto*-de re distinction to propositional attitude constructions can, however, lead to surprising, even paradoxical, results. Here I am going to present a version of an argument due to Alonzo Church for the conclusion that it is not possible to conceive of one and the same thing as being two. If Church's argument is correct it seems to threaten the new wave position and perhaps bolster a property-dualistic position.

6.4. Church's paradoxical argument:

Adapting an argument due to Alonzo Church (1988) to the case of de re imaginability, we can prove from intuitively reasonable premises that:

(I) $\forall x \forall y(\text{Imagines}(\text{John}, x \neq y) \rightarrow x \neq y)$.

We can even strengthen the formulation of (I) to:

(II) $\forall x \forall y(\text{Imaginable}(x \neq y) \rightarrow x \neq y))$,

that is, if it is imaginable that x and y are distinct, then they are distinct.

With respect to the proof below, it does not matter whether we speak of imaginability or conceivability. So the following principle for conceivability is also forthcoming:

---

[10] Cf. Burge (1977), Kaplan (1969), and Kaplan (1986).

(III) $\forall x \forall y(\text{Conceivable}(x \neq y) \rightarrow x \neq y)$.

We give the proof for the case of conceivability:

First we have the logical principle of *indiscernibility of identicals* (or *Leibniz' Law*):

(1) $\forall x \forall y(x = y \rightarrow (\varphi(x) \rightarrow \varphi(y)))$.

From this we get:

(2) $\forall x \forall y(x = y \rightarrow (\text{Conceivable}(x \neq y) \rightarrow \text{Conceivable}(y \neq y))$.

The following principle appears to be *a priori* and self-evident:

(IV) $\forall x \neg \text{Conceivable}(x \neq x)$,

i.e., it is not conceivable of an object that it is distinct from itself.
From (2) and (IV) we get by elementary logic:

(3) $\forall x \forall y(x = y \rightarrow \neg \text{Conceivable}(x \neq y))$.

But this is logically equivalent to (III). Q.E.D.

Suppose now that we have two singular terms **a** and **b** of the kind that is often referred to as "Millian names", or "directly referential" singular terms. This means that the sole semantical contribution of one of these terms, in a given context of use, to the proposition expressed by any sentence in which the term occurs, is just the referent of the term.

But if the sole semantic contribution of a term is its referent, it is hard to see how the substitution of coreferential such terms could change the truth value of any (non-quotational) sentence in which it occurs.

It is also hard to see how the principles of existential generalization and universal instantiation could fail for such terms, even in modal contexts or in propositional attitude contexts. We should therefore be justified in inferring from (III):

(III') $\text{Conceivable}(\mathbf{a} \neq \mathbf{b}) \rightarrow \mathbf{a} \neq \mathbf{b}$,

provided **a** and **b** are Millian, or directly referential, names.

So if 'Phosphorus' and 'Hesperus' are two Millian names of the planet Venus, we should have:

(1) $\text{Conceivable}(\text{Phosphorus} \neq \text{Hesperus}) \rightarrow \text{Phosphorus} \neq \text{Hesperus}$.
(2) $\text{Phosphorus} = \text{Hesperus}$.

Hence,

(3) $\neg \text{Conceivable}(\text{Phosphorus} \neq \text{Hesperus})$.

However, this conclusion is quite unintuitive.

6.5. The mental and the physical

Let us consider the following piece of reasoning:

(1)	Conceivable(Pain ≠ ECF)	Premise
(III)	$\forall x \forall y$(Conceivable(x ≠ y) → x ≠ y)	
(2)	Conceivable(Pain ≠ ECF) → Pain ≠ ECF	from (III) by U.S.
(3)	Pain ≠ ECF	from (1), (2) by M.P.

That is, if it is possible to conceive of Pain being distinct from C-fiber excitation, then Pain *is* distinct from C-fiber excitation. Isn't this argument too good to be valid? We didn't even use the controversial principle that conceivability implies possibility.

Wait a moment. Couldn't we use the same kind of argument to prove that Water is distinct from H_2O? Surely, we can conceive of Water as being something else than H_2O. But Water, we know, *is* H_2O. So, we have a contradiction. How do we get out of this conundrum?

The assumption that proper names like 'Phosphorus' and 'Hesperus', 'Pain', 'ECF', 'Water', and 'H_2O', are Millian, or directly referential, names is stronger than the assumption that they are "rigid designators". Usually it is assumed that directly referential terms are also rigid designators, but the converse obviously does not hold.

'Water' and 'H_2O' may be rigid designators as Kripke claims, but they are hardly directly referential, in the sense of lacking descriptive meaning. But this means that the inference from:

(1) Conceivable(Water ≠ H_2O)

to

(2) Water ≠ H_2O,

does not seem to be valid.

The statement (1) is *de dicto* rather than *de re*. For the same reason, the inference from Conceivable(Pain ≠ ECF) to Pain ≠ ECF can fail.

So our way of inferring Pain ≠ ECF from Conceivable(Pain ≠ ECF) does not go through.

Kripke's argument (Se Rabinowicz's contribution to this volume) has the apparent advantage that it only requires that the terms 'Pain' and 'ECF' are

rigid designators, not that they are directly referential. On the other hand it uses the questionable bridge-principle from conceivability to possibility.

But perhaps we can still save the argument from the conceivability of Pain ≠ ECF to the conclusion that Pain ≠ ECF. Suppose that we change the premise of the argument to:

(1) $\exists x \exists y(x = \text{Pain} \wedge y = \text{ECF} \wedge \text{Conceivable}(x \neq y))$.

From this premise together with:

(III) $\forall x \forall y(\text{Conceivable}(x \neq y) \rightarrow x \neq y)$

we can infer:

(2) $\exists x \exists y(x = \text{Pain} \wedge y = \text{ECF} \wedge x \neq y)$

from which follows:

(3) Pain ≠ ECF.

Then, we would have an argument that seems to establish the desired conclusion.

How could we establish the premise (1)? Someone might suggest that this is easy. Just concentrate on a particular kind of pain, when you have it, and give it a name **a**. Then name the correlated brain state – that you observe in your Autocerebroscope – **b**.

Naming one's own mental states is certainly not an easy task, not to speak of naming determinate brain states. But even if you succeeded, could you be sure that it is conceivable that **a** ≠ **b**? Someone antecedently convinced of **a** = **b** would presumably not be impressed, and question the conceivability of **a** ≠ **b**. He might suggest that you have misidentified the mental state **a** or the corresponding physical state **b**, or both.

Some philosophers maintain that we have special so-called *phenomenal concepts* available – in our language of thought, so to speak – by means of which we can directly refer to our own mental states. These concepts could perhaps then play the role of **a** in the argument. The difficulty remains, however, of finding a term **b** that could serve as logical proper name of the brain state that is the neurological correlate of the mental state **a**. Alternatively, we would need an argument establishing:

$\forall y(\text{Brain State}(y) \rightarrow \text{Conceivable}(\mathbf{a} \neq y)$.

Then, we could use (III) to establish:

$\forall y(\text{Brain State}(y) \rightarrow \mathbf{a} \neq y)$,

for any phenomenal state **a**. So perhaps our principle (III) is not completely useless. However, in the next section I am going to put (III) in question. Perhaps the proof that we gave for (III) can be questioned after all.

6.6. A second look at the paradoxical argument

In order to discuss the argument that was presented above for the principle:

(III) $\forall x \forall y(\text{Conceivable}(x \neq y) \rightarrow x \neq y)$,

I am going to introduce an alternative notation to describe *de re* propositional attitudes that was introduced by Quine in a classical paper (Quine 1956). Instead of analyzing statements like:

(1) John imagines of the substances Water and H_2O that they are distinct,

by means of "quantifying in"-constructions:

(2) $\exists x \exists y(x = \text{Water} \wedge y = H_2O \wedge \text{Imagines}(\text{John}, x \neq y))$,

Quine uses a primitive notation for *de re* attitudes:

(3) $\text{Imagines}(\text{John}, xy[x \neq y])(\text{Water}, H_2O)$.

This should be read:

(4) John imagines the relation $xy[x \neq y]$ true of (Water, H_2O).

Thus, we can think of "imagines" as an intensional operator which given an individual term t and a term designating an n-ary relation in intension P, yields a new n-ary relation:

(5) Imagines(t, P).

By means of (5), we can form *de re* statements like:

(6) $\text{Imagines}(t, P)(t_1,..., t_n)$,

with the intended meaning that t imagines the relation P true of $(t_1,..., t_n)$.

In addition, there is an abstraction operator $x_1...x_n[...]$, which given any formula $\varphi(x_1,..., x_n)$ with distinct free variables $x_1,..., x_n$ forms a name $x_1...x_n[\varphi(x_1,..., x_n)]$ of the relation (in intension) determined by $\varphi(x_1,..., x_n)$. Intuitively, $x_1...x_n[\varphi(x_1,..., x_n)](t_1,..., t_n)$ says that $t_1,..., t_n$ (taken in that order) stand in the relation defined by $\varphi(x_1,..., x_n)$ to each other. Thus, it holds that:

(7) $\forall x_1...x_n \mathbf{L}(x_1...x_n[\varphi(x_1,..., x_n)](t_1,..., t_n) \leftrightarrow \varphi(t_1,..., t_n))$.

Using Quine's notation, we can distinguish between:

(10) Conceives(John, x[x ≠ x])(Venus)

and

(11) Conceives(John, xy[x ≠ y])(Venus, Venus).

Intuitively, (10) says that John conceives of Venus as having the property of self-distinctness. That is, John conceives of Venus as having a contradictory property. (11), on the other hand, says that John conceives of Venus as having a certain relation to itself, namely the relation of distinctness.

The distinction between (10) and (11) may seem minute. But suppose that John has two distinct proper names for Venus, say 'Phosphorus' and 'Hesperus', and thinks that the following statement may describe a true state of affairs:

(12) Phosphorus ≠ Hesperus.

Perhaps, this is sufficient for the truth of:

(13) Conceives(John, xy[x ≠ y])(Phosphorus, Hesperus).

Since Phosphorus = Hesperus = Venus, the following is then also true:

(14) Conceives(John, xy[x ≠ y])(Venus, Venus).

But

(15) Conceives(John, x[x ≠ x])(Venus)

does not hold in the described situation, since John would never assent to a statement of the form t ≠ t, for instance, Phosphorus ≠ Phosphorus.

In ordinary quantified predicate logic, the distinction between (14) and (15) collapses. They become, respectively:

(16) $\exists x(x = \text{Venus} \wedge \text{Conceives(John, } x \neq x))$

(17) $\exists x \exists y(x = \text{Venus} \wedge y = \text{Venus} \wedge \text{Conceives(John, } x \neq y))$,

which are logically equivalent.

Let us now try to reproduce the proof of:

(III) $\forall x \forall y(\text{Conceivable}(x \neq y) \rightarrow x \neq y)$

using Quine's notation. Thus, we want to see whether we can prove:

(V) $\forall x \forall y(\text{Conceivable}(x,y[x \neq y])(x, y) \rightarrow x \neq y)$.

Intuitively this should not be possible given the interpretation of *de re* conceivability that we are contemplating.

We could have two *de re* representations 'Phosphorus' and 'Hesperus' *of* the same object Venus, without realizing that they are of the same object and

therefore conceiving of the statement 'Phosphorus ≠ Hesperus' as being possibly true. So even if assume:

(1) $\forall x[\neg \text{Conceivable}(x[x \neq x])(x)]$

we should not be able to prove (V).

If, however, we could derive:

(2) $\forall x[\text{Conceivable}(xy[x \neq y])(x, x)]$

from (1), we could proceed to (V) from (2). Namely, we have:

(3) $\forall x \forall y[x = y \rightarrow \text{Conceivable}(xy[x \neq y])(x, y) \rightarrow \text{Conceivable}(xy[x \neq y])(y, y)]$.

So, from (2) and (3) we would get:

(4) $\forall x \forall y[x = y \rightarrow \neg \text{Conceivable}(xy[x \neq y])(x, y)]$,

which is equivalent to (V).

But (2) does not follow from (1) on the suggested interpretation of *de re* conceivability. An informal argument that seems to be valid when formalized in a standard quantified intensional logic turns out not to be valid when formalized in another logical language, a language that is able to make finer conceptual distinctions. Quine's skepticism towards quantified intensional logic and "quantifying in" does not seem to have been entirely unfounded.

The moral I think we can draw from this exercise is that there is no logical reason to accept the principle:

If we can conceive of two entities as being distinct then they are distinct.

This opens up the possibility for a physicalistic theory of mind according to which mental states are identical with physical states although it is conceivable that they may not be so identical. Of course there could still be philosophical arguments against such a view. But I think it is unlikely that such arguments should be based on conceivability considerations.

References

Almog, J., 2002, *What am I? – Descartes and the Mind-Body Problem,* Oxford University Press.

Barcan Marcus, R., 1983, 'Rationality and Believing the Impossible', *Journal of Philosophy*, LXXX, 6, pp. 321-338. Reprinted in Barcan Marcus (1993).

Barcan Marcus, R., 1993, *Modalities - Philosophical Essays*. Oxford University Press.

Burge, T., 1977, 'Belief De Re', *Journal of Philosophy* **74**, 338-362.

Church, A.1988, 'A Remark Concerning Quine's Paradox about Modality', in Salmon, N. and Soames, S., *Propositions and Attitudes*, Oxford: Oxford University Press.

Davidson, D., 2001, 'The Emergence of Thought', in Davidson, D., *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press.

Fagin, R. and Halpern, J. Y., 1988, 'Belief, Awareness and Limited Reasoning', Artificial Intelligence, 34, pp. 39-76.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y., 1975, *Reasoning about Knowledge*, Cambridge, Massachsetts: The MIT Press.

Gettier, E., 1963, 'Is Justified True Belief Knowledge?' Analysis, 1963, pp. 121-123.

Hintikka, J., 1962, *Knowledge and Belief,* Cornell University Press, Ithaca.

Hintikka, J., 1969, *Models for Modalities,* D. Reidel, Dordrecht.

Horgan, T. and Tienson J., 2001, 'Deconstructing New Wave Materialism', in Gillett C. and Loewer, B. (eds) *Physicalism and Its Discontents*, Cambridge: Cambridge University Press.

Kaplan, D., 1969, 'Quantifying In', in D. Davidson and J. Hintikka (eds.), *Words and Objections: Essays on the Work of W. V. Quine,* D. Reidel, Dordrecht, pp. 178-214. Reprinted in Linsky (1971).

Kaplan, D., 1986, 'Opacity', in *The Philosophy of W. V. Quine*, The Library of Living Philosophers Volume XVIII, edited by Lewis Edwin Hahn and Paul Arthur Schilpp, pp. 229-294. Open Court, Chicago.

Kripke, S., 1979, 'A Puzzle about Belief', in A. Margalit (ed.), *Meaning and Use,* D. Reidel, Dordrecht, 239-283.

Kripke, S., 1980, *Naming and Necessity,* Basil Blackwell, Oxford.

Linsky, L. (ed.), 1971, *Reference and Modality,* Oxford University Press, London.

Meyer, J.-J. Ch., 2001, 'Epistemic Logic' in Goble, L. (ed.) *The Blackwell Guide to Philosophical Logic*, Oxford: Blackwell.

Putnam, H. (1990). 'Is water necessarily H_2O?', Chapter 4 in *Realism with a Human Face*, (pp. 55–79). Cambridge, MA: Harvard University Press.

Quine, W. V., 1956, 'Quantifiers and Propositional Attitudes' *Journal of Philosophy* **53**, 177-187. Reprinted in Linsky (1971).

Sorensen, R., 1996, 'Modal Bloopers: Why Believable Impossibilities are Necessary', *American Philosophical Quarterly*, 33/1, pp. 247-261.

Stenius, E., 1963, *Wittgenstein's 'Tractatus' – A Critical Expositions of the Main Lines of Thought*. Oxford: Basil Blackwell.

Wittgenstein, L., 1922, *Tractatus Logico-Philosophicus*. London.

Yablo, S, 1993, 'Is Conceivability a Guide to Possibility?', *Philosophy and Phenomenological Research* Vol LIII, No. 1, pp. 1-41.