Ex Nihilo, Volume VII
The Undergraduate Philosophy Journal of
The University of Texas at Austin
Fall 2008

Cover Design
Julia Bond
Ryan Haecker

Cover Artist
Gustav Doré, *The New Zealander*

Ex Nihilo, Volume VII
The Undergraduate Philosophy Journal of
The University of Texas at Austin
Fall 2008

Contents

Essays

Editorial Foreword

As undergraduates, the opportunity not only to publish intellectual work in a journal, but also to engage a discourse community made up of our peers from universities across the globe is of paramount importance to us, the editors of this publication, and to you, the authors and readers who make Ex Nihilo possible. From the call for papers that prompts students from all over the country to submit their best work, through the intense editorial review process, to the annual conference where the authors of exceptional articles come to our university for intellectual fellowship; every step of the process of creating this journal is an invaluable one on the path towards becoming better philosophers and students.

Of all of the aforementioned steps of the process, that of selecting the essays that ultimately find their place between these covers is without question the most difficult. Evaluating papers from a self-selected group of high achievers in thought and writing is extremely difficult, and we received many more meritorious papers than we are able to publish. We selected the papers that we believe best met our criteria of significance, rigor, and clarity.

Accordingly, a word or two on our selection process is in order. This year, we divided up the submissions into several high level categories according to their subject matter. The categories we applied included: language and mind, metaphysics and epistemology, history of philosophy, philosophy of religion, and continental philosophy. Groups of three members of the editorial board were then assigned, according to their interests and training, to review the essays in each category. After reviewing the essays, each group met in person to argue for or against advancement of a particular essay to the next and final tier of the review process. Those essays that made it to the final review were then carefully read and evaluated by all members of the editorial board before being finally and democratically accepted or rejected.

The contents of the journal before you are the result of this long and difficult process, a process that we hope has led to the production of an undergraduate journal that will reward its readers as much as it has rewarded us.

Coeditors

Zombies Are Here to Stay[1]

Daniel Kirksey
Kansas State University

McLaughlin and Hill[2] reject the anti-materialist argument made by Chalmers[3]. By their reconstruction Chalmers' argument takes the following form:

1. **In our world, there are conscious experiences.**

2. **There is a logically possible world physically identical to ours [i.e., there is a physical duplicate of our world] in which the positive facts about consciousness in our world do not hold.**[4]

3. **Therefore, facts about consciousness are further facts about our world, over and above the physical facts.**

4. **So materialism is false.**

McLaughlin and Hill challenge the soundness of this argument by rejecting premise two in virtue of the argument upon which it rests; more specifically, they reject the conceivability-possibility argument as invalid in its application to psychophysical identities. Ultimately, I intend to show that McLaughlin and Hill's analysis is misapplied for two reasons. First, their analysis of the conceivability-possibility argument, even within their explicitly limited scope of its application to psychophysical identities, does not lead to a rejection of the principle as they claim. Secondly, if handled differently,

[1] I would like to thank Dr. Marcelo Sabatés for helping me form an understanding of the mind-body problem.

[2] Christopher S. Hill and Brian P. McLaughlin, *There are Fewer things in Reality than are Dreamt of in Chalmers's Philosophy* (Philosophy and Phenomenological Research Vol. LIX, No. 2, 1999). Hereafter, all parenthetical page references are to this volume.

[3] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford and New York: Oxford University Press, 1996). Hereafter, all parenthetical page references are to this volume.

[4] This premise is based on the well-known "zombie" thought experiments.

what their analysis can do is allow us to see why the conceivability-possibility principle is irrelevant in the question of materialism vis-à-vis property dualism.

I. The Method and Intention

The question provoking McLaughlin and Hill's response to Chalmers is to how he shows premise two to be true. The principle, on which the truth of premise two rests, and what McLaughlin and Hill deny, is the conceivability-possibility argument implicit in the conceivability-possibility principle. The invocation of this argument can be seen when Chalmers states that one can imagine a possible world with the identical positive physical facts of our world at time T_1 holding, but without the positive facts of consciousness at T_1 in our world holding (Chalmers 131). The conceivability-possibility principle can be formalized as:

C: ***"is conceivable"*** **P:** ***"is logically possible"***

(x) (Cx → Px)

This principle – that anything conceivable is at least logically possible – is what McLaughlin and Hill reject as invalid. The reason they give for doing so is problematic.

McLaughlin and Hill's rejection of the conceivability-possibility argument is delineated in their explanation of what they call the *a posterioriocity* of psychophysical identities[5]. Embedded in this explanation is their argument against the validity of the conceivability-possibility argument utilized by Chalmers. McLaughlin and Hill's argument, though not explicitly given, is structured in the following way:

1. **We can conceive of a logically possible world W, physically identical to ours, in which the positive facts about consciousness in our world do not hold.**

2. **The process in which we conceive of W (with its lack of consciousness) has certain flaws F.**

3. **Concerning psychophysical identity, conceivability does not imply possibility.**

4. **Premise 2 of Chalmers' argument is true just in case conceivability implies possibility.**

5. **Therefore, Chalmers' argument is unsound.**

[5] Reminiscent of Kripke's burden of responsibility to explain why the identity of H_2O = water is not contingent (meaning that names "H_2O" and "water" express rigid designators), irrespective of the fact that the relationship can only be known a posteriori, McLaughlin and Hill inherit the same responsibility in their claim of necessary psychophysical identity that can only be known a posteriori.

One's attention is immediately drawn to the jump from premise two to premise three. This step is given much attention by McLaughlin and Hill and necessarily so, as it is not clear how it follows. More needs to be said about F, if it is the case that premise three follows.

I now turn to McLaughlin and Hill's discussion of F in their hope of closing the gap between premises two and three. They start by making a distinction between types of conceptualizing. To *imagine*, is "…to try to construct a qualitative or imagistic representation…" of a given situation – the given situation is an attempt at creating a counterexample to the conceivability-possibility argument as it is concerned with psychophysical identities (McLaughlin and Hill 447). The second type of conceptualization is *conceiving*, i.e., "…try[ing] to construct a conceptual representation of [the same counterexample attempted in *imagining*] – a representation that has a logical structure and that has concepts as its basic building blocks" (McLaughlin and Hill 447). The distinction between *imagination* and *conceptualization* appears to be nothing more than the presence of mental images in imagination and the lack thereof in conceiving. Though McLaughlin and Hill found the distinction to be of enough importance to discuss it in detail, I am under the impression that the distinction between *imagining* and *conceptualizing* is not necessary in order for them to reach their conclusion[6]. With that said, I will focus on the argument built around the imagination form of conception, while appealing to conceiving as *conceptualizing* insofar as it serves to corroborate their point.

McLaughlin and Hill's analysis of the mode in which a person conceptualizes in the imaginative sense brings another distinction between what they call *sympathetic* and *perceptual* imagination, and it is in this distinction that we find the force and fallaciousness of their argument. They credit Nagel as the first to make the distinction in his article "What Is It Like to Be a Bat" saying:

> *"…To imagine something perceptually, we put ourselves in a conscious state resembling the state we would be in if we perceived it. To imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. (This method can be used to imagine mental events and states – our own or another's.) When we try to imagine a mental state occurring without its associated brain state, we first sympathetically imagine the occurrence of the mental state: that is, we put ourselves in a state that resembles it mentally. At the same time, we attempt to perceptually imagine the non-occurrence of the associated physical state, by putting ourselves into another state unconnected with the first: one resembling that which we would be in if we perceived the non-occurrence of the physical state." (442)*

Nagel is describing the process by which human beings conceptualize – what he asserts (as McLaughlin and Hill agree) is that there is a kind of self-deception in the thought process. The much-used "zombie" thought experiment is often used to make Nagel's point more concrete. In the thought experiment a person first imagines what it is like to be conscious, i.e., imagine what it's like to have mental states (or not to have them). This is what Nagel calls the sympathetic stage; it is an experience. The second stage in this

[6] I thank Brian McLaughlin for his advice about the differences between *imagining* and *conceptualizing*, in which, he elaborated upon the relationship between the different distinctions and their equal conclusions.

thought experiment is the point of confusion. While imagining what it's like to have mental states (or not to have them), one simultaneously imagines the absence (or presence) of the associated brain state. This image is the perceptual stage; it is an observation. A person doing this has switched from sympathetic imagination to perceptual, i.e., he is no longer experiencing, but observing according to Nagel. The difference between experiencing and observing is what McLaughlin and Hill seem to mean in their discussion of conceptualizing using concepts. They say:

> *"When one uses a sensory concept to classify one's own current experiences, the experiences that guide and justify one in applying the concept are always identical with the experiences to which the concept is applied. Sensory states are self-presenting states: we experience them, but we do not have sensory experiences of them. We experience them simply by virtue of being in them…An agent's access to the phenomena that he or she perceives is always indirect: it always occurs via an experience of the perceived phenomena that is not identical with the perceived phenomena, but rather caused by it." (448)*

When one imagines a mental state, he is simply experiencing a mental state. Imagining a mental state is identical to being in a mental state. There is no difference between existing as a conscious being and imagining a mental state. We do not experience the experience of a mental state or experience the experience of an experience. In this sense, imagining is being. However, when an agent is imagining in the perceptual sense, there is a difference between the experience and that which is causing the experience: they are two distinct entities. In the perceptual sense, imagination needs something outside of experience – something more than itself. From this McLaughlin and Hill conclude that it is possible to conceive of a world in which there are mental states that have no corresponding brain states. However, a close analysis of the process by which one is said to conceive of such a world reveals an incoherency whereas a person equivocates the two distinct entities of experience and observation – albeit perhaps unknowingly.

II. McLaughlin and Hill's Equivocation

Recall that McLaughlin and Hill's intention is to reject premise two of Chalmers' argument by denying the validity of the argument upon which it rests:

> **[p1] One can imagine all the physical facts holding without the facts about consciousness holding, [c] therefore, the physical facts do not exhaust all the facts. (McLaughlin and Hill 445)**

They present an apparent counterexample in which p1 is true, but c does not follow. The question is then, whether their explanation of how we conceive of a world physically identical to our own without identical states of consciousness, allows a justified claim of **$\sim(x)(Cx \rightarrow Px)$**, i.e., have they shown an occurrence of **$Ca \wedge \sim Pa$**?

Consistent with my first claim about the overall misapplication of their analysis – no, McLaughlin and Hill have not shown such an occurrence specifically because they have not shown an occurrence of **Ca**. Succinctly stated, McLaughlin and Hill have equivocated the meaning of "conceive of" and what I will describe as "think about." Making this distinction significantly damages McLaughlin and Hill's claims against Chalmers' second premise. As I will attempt to show, I think this distinction must be made.

In order to see the merit of what I maintain to be a clear difference in meanings between "conceiving of" and "thinking about," it is necessary to take a closer look at what compels my intuition. From McLaughlin and Hill's zombie analysis, we can infer their beliefs as to what constitutes sufficient conditions in order to claim one is conceiving of a given something. As far as I can tell, the only requirements to be inferred from McLaughlin and Hill are non-existent. Recall the case at hand, in which, McLaughlin and Hill claim a person is conceiving of a possible world that is logically impossible. The strongest condition I can infer from this claim is that conceivability requires mental activity. There is no need for logical coherency or even consistency. For McLaughlin and Hill, as long as there is an occurrence of some cognitive activity at time T_1, there is an instance of conceiving at time T_1, which is to say nothing about what it means to conceive of something.

The difference between what McLaughlin and Hill deem *conceiving* and what I call *thinking about,* can be seen more clearly when considering the following problem:

$\perp$ = a logical contradiction.

S1: We can conceive of $\perp \rightarrow \perp$ is possible

I've formed this conditional as such because there is no question or debate about the logical possibility of the consequent (i.e. lack thereof), or the absurdity entailed by accepting the entire conditional. The sufficient conditions that constitute what it is to conceive, as implied by McLaughlin and Hill discussed earlier, leave us no reason to believe that one cannot conceive of $\perp$. But, what choices do we have in the matter at this point. Unfortunately for McLaughlin and Hill, there are no attractive choices. They have two: they must give up the idea that one can conceive of $\perp$, which will provoke a new set of necessary conditions that must be met in order to deem any given mental activity as *conceiving*, and therefore, undermine their argument against Chalmers. Or, they can accept the entire conditional and the absurdity that comes with it.

Notice that the same problem does not exist when substituting *thinking about* for *conceiving* in the following conditional:

S2: We can think about $\perp \rightarrow \perp$ is possible

Like **S1**, there is no question about the falsity of the consequent or the absurdity of accepting the entire conditional. The only possible option is to accept the antecedent; that is to say, allow that one can be said to be thinking about a logical contradiction, while retaining the claim that logical contradictions are logically impossible.

To say one can *conceive* of something requires a certain level of coherency that is not required when one is said to be thinking about something. One can *think about* a square circle, and even be completely convinced that, for at least a moment, a single image existed in the mind that simultaneously satisfied everything that it is to be a circle, and everything that it is to be a square. But, there is no reason to doubt that a close analysis of what McLaughlin and Hill would call the process of conceiving of a square circle, would reveal the same kind of incoherency they found in the zombie case. There is a reason why McLaughlin and Hill have not been arguing against a *thinking about-possibility* argument – it's because of the self-evident fact that certain kinds of mental activity (e.g. misunderstanding, confusion, incoherent, streaming consciousness, uninformed, etc.) are obviously unreliable guides to truth and possibility. The zombie thought experiment is a clear case of this kind of mental activity.

III. The Inevitable Failure

McLaughlin and Hill's equivocation results in a failure to refute Chalmers' argument, which rests on the conceivability-possibility conditional. Recall premise two and three of my reconstruction of their main argument:

2. **The process in which we conceive of W has certain flaws F.**

3. **Concerning psychophysical identity, conceivability does not imply possibility.**

I stated that there was a gap between premise two and premise three in need of an explanation for the validity of the entire argument to follow. McLaughlin and Hill focus on premise two, and show how an agent's conception can be flawed in a way that does not allow us to say that the agent is conceptualizing something that is actually possible. That being the case, their argument's validity problem based on the jump from premise two to premise three, becomes a soundness problem concerning their first premise:

1. **We can conceive of a logically possible world W, physically identical to ours, in which the positive facts about consciousness in our world do not hold.**

McLaughlin and Hill accept the antecedent of the conceivability-possibility conditional upon which premise two of Chalmers' argument rests, which is their premise one; however, they do not accept the conclusion. Their error is the acceptance of the antecedent on the grounds of Nagel's description of the way in which we can be said to be conceiving of the zombie world – bringing the problems discussed in the previous section.

IV. Toward A Better Application

McLaughlin and Hill claim that we can only know psychophysical identities a posteriori because of the problems described in Nagel's passage and I agree. However, given the analysis of the brain processes used for the zombie thought experiment by Nagel, it seems like a focus on premise two of Chalmers' argument is misdirected in any effort to deny the conceivability-possibility principal upon which it rests. This is true for both materialists and property dualists. We cannot conceive of the zombie world because of what seems to be epistemic constraints due to the way in which our minds work – there is no reason to think that these constrains are exclusive to either a materialist world, or a dualist world. Given that, if we cannot conceive of the possible world in question because of the reasons given by Nagel, then we can't take McLaughlin and Hill's main argument as especially useful. With that being the case, the same must be said for Chalmers' argument as it has been presented in this article. With that said, the conceivability-possibility principle is at this point irrelevant to the mind body problem.

Daniel Kirksey
Kansas State University

References

Chalmers, David. The Conscious Mind: In Search of a Fundamental Theory (Oxford and New York: Oxford University Press, 1996).

Hill, Christopher S. and McLaughlin, Brian P. "There are Fewer things in Reality than are Dreamt of in Chalmers's Philosophy,'' Philosophy and Phenomenological Research 59, No. 2 (1999).

Nagel, Thomas. "What Is It Like to Be a Bat?," The Philosophical Review 83 (1974): 435-50.

What Does the Liar Say About the Truth?

Simona Aimar
St. Andrews University

ABSTRACT

Many modern approaches to the Liar paradox have abandoned the search for its conclusive solution. The paradox is taken to have a lesson to teach about the notion of truth. Alfred Tarski famously drew from the liar the need of a hierarchy of indexed truth predicates, in order to achieve consistency. Saul Kripke, in the seventies, proposed an alternative for a consistent theory of truth, by partially defining a single truth-predicate. Kripke's solution is generally regarded as scoring better than Tarski's, in that it allows for self-referential claims. Together with subsequent formal theories of truth, Kripke and Tarski's proposal share a charge of counter-intuitivity. Toy-models for a language, it is argued, reach consistency by introducing some restriction on the truth-predicate. By so doing, however, they unduly depart from our intuitive concept of truth. The paper defends certain semantic models against such objection. Building on a suggestion by Colin Howson, it is argued that Kripke's and Jon Barwise and John Etchemendy's proposals can handle the counter-intuitivity charge, if they are read as proposing the adoption of two *complementary* truth-notions.

The liar paradox, in one of its simplest versions, is the following. Consider the sentence

(L) this sentence is false,

where "this sentence" refers to (L). Suppose (L) is true. Given that (L) is true iff what it says is the case, then what it says is the case. (L) says of itself that it is false. Hence, L is false. Contradiction. Conclude, by negation introduction, that (L) is false. Given that (L) is false iff what it says is not the case, then what it says is not the case. (L) says of itself that it is false. Hence, reading "not false" as "true", (L) is true. Paradox.

Most of modern approaches to the liar paradox have abandoned the search for its conclusive solution. Investigations of the logic of truth take the paradox as a challenge for any definitional attempt of the truth-notion. The stream starts off with Alfred Tarski. In 1933, Tarski made of liar-like contradictions a *reductio* proof: no language that follows the laws of classical logic, he contended, can consistently contain its own truth-predicate – i.e. be semantically closed.[1] Tarski proposed to define truth by means of a series of typed truth-predicates, thereby developing his well-known hierarchy of semantically open object languages, meta-languages, meta-meta-languages etc. Saul Kripke and the team of Robert Martin and Peter Woodruff, in the mid 1970s, extended Tarski's methods to semantically closed languages. The liar paradox is here avoided by defining the truth-predicate as partial; some sentences are assigned to the gap between truth and falsity.[2] What follows is a proliferation of models for semantically closed languages. John Barwise and John Etchemendy, for instance, combine non well-founded set theory, an Austinian account of propositions and situation semantics. The theory they come up with considers truth as situation dependent and liar propositions as not paradoxical.[3]

Their popularity notwithstanding, formal theories of truth share a standard objection: the charge being of an excessive departure from our intuitions. Graham Priest, among others, has observed that toy-models for a language do not give us an analysis of the truth-notion as deployed in the original liar reasoning. Intuitively, truth is understood holistically. Semantic theories such as those mentioned above, however, reach consistency by introducing restrictions on the truth-predicate. All these theories, moreover, are stated and appraised on the backdrop of a broader truth-notion, often confined to a meta-language level. This seems to show that the solution they adopt in order to avoid the liar paradox is unsatisfactory.[4]

In this essay, I shall defend certain semantic models against the objection of counter-intuitivity. I shall do so in two steps. First, I shall suggest that Tarski's theory of truth contains important insights about the truth-notion. Tarski's proposal does fall prey to fatal counter-intuitivity objections. However, one might regard some of its crucial

[1] Tarski: 1933.

[2] Kripke: 1975, Martin and Woodruff: 1975.

[3] Barwise and Etchemendy: 1987.

[4] Cf. Priest: 1987a, 1987b and 1993. Mentioning the so-called "Liar Revenge" often makes the same point. For all the diagnoses of the liar paradox seen above, a pathological consequence seems to follow: they are unable to accommodate all intuitively true claims by means of the very truth-predicate(s) thereby defined.

features as interesting insights about the truth-notion. Second, I shall argue that subsequent theories of truth can better handle the counter-intuitivity charge. One of the main reasons for which, say, Kripke's appeal to a meta-language is seen as unsatisfactory is that its relation with our intuitions is left unexplained. Building on a suggestion given by Colin Howson, I shall argue that Kripke's and Barwise and Etchemendy's theories can be interpreted as offering such an explanation. They might be read as showing that our holistic conception of truth is *complementary* to the partial truth-notion they define.[5]

In section I, I will outline and appraise Tarski's view on the liar paradox.
Section II will be devoted to critical discussion of Kripke's theory of truth. Section III will focus on Barwise and Etchemendy's model.

1.1 Tarski. Truth, Undefinability and Hierarchies

The formal study of the truth-notion begins with Tarski's T-schema.[6] On the stipulative assumption that truth is a predicate for sentences, Tarski contends that a truth-predicate T(x) for a language *L* is adequately defined if the following schema holds

(T-schema) T(φ) iff φ,

where φ stands for any well-formed sentence of *L*. The equivalence indicates as a necessary condition for defining the truth-predicate a series of quasi-trivial biconditionals. This is why it looks very plausible. Nevertheless, the schema allows, under certain conditions, for inconsistencies. Given a sentence φ* that expresses its own falsity (henceforth, the Liar sentence), if we read "false" as "not true" by T-schema we obtain

T(φ*) iff not-T(φ*),

which is a plain contradiction.

As is well-known, on Tarski's view the liar contradiction represents a *reductio ad absurdum*, to be read as indicating the further conditions that a consistent theory of truth has to satisfy. The liar paradox is thereby turned into the following theorem:

> (Undefinability Theorem) No consistent language in which the ordinary rules of logic hold can be semantically closed.

The theorem is supposed to sum up the liar's lesson. Lest abandoning classical logic, it claims, any language containing its expressions, the denotation terms for those expressions and its own truth-predicate (that is, any semantically closed language) cannot be consistently formalized. On the face of it, Tarski draws two further conclusions. First, he takes the theorem as establishing the inconsistency of a natural language like English, this being complex enough to achieve semantic self-reference. Second, he addresses his search for a consistent theory of truth to formal languages only.[7]

[5] Howson: forthcoming.
[6] Tarski: 1933 and 1944.
[7] Tarski: 1933 and 1944.

The theory of truth Tarski comes up with banishes any application of the truth-predicate to sentences containing it. A truth-predicate is defined by considering more than one language at once. On the one hand, we take an object language *L*, whose sentences we want to define a truth-predicate for. On the other hand, we introduce a meta-language *M*, to which the truth-predicate belongs. This second language, as the original object language, is not semantically closed. Its own semantic properties are predicated by a further meta-meta-language, and so on. One can thus build a consistent hierarchy of semantically open languages. At the bottom level, there is no sentence containing the truth-predicate (or related terms). At each further level, there is a distinct truth-predicate that can only be applied to the sentences of a lower level. Since the truth-predicate involved in the liar sentence needs to refer to a sentence of the same level, the construction is prevented from the liar paradox.

1.2 An evaluation of Tarski's proposal: counter-intuitivity and interesting remarks.

Tarski's hierarchy successfully avoids inconsistencies. By restricting the domain of each truth-predicate to the sentences of another language, no liar sentence is well formed; *a fortiori*, no liar paradox can arise. Nevertheless, the theory has been strongly criticized. The main charges it has been hostage of emphasize its counter-intuitivity. Here, I shall consider two versions of the counter-intuitivity objection. I shall try to suggest that a rebuttal of Tarski's theory as utterly inadequate might be too harsh an assessment to draw.

The first objection has it that Tarski's proposal dismisses natural languages as incoherent too quickly. As briefly mentioned above, Tarski strongly doubts that the search for a consistent definition of truth should be extended to semantically closed languages.[8] The results of such defeatism, however, are quite drastic. Tarski ends up declaring a definition of our ordinary truth-predicate an impossible task. This, in addition, seems to entail a dismissal of the ordinary truth-concept. For that concept underlies, say, the English expression 'is true'. A second objection addresses Tarski's theory insofar as directed towards formal languages. The charge concerns Tarski's fragmentation of the truth-notion. There seems to be something unacceptable in the idea of adopting a series of language-indexed truth-predicates. Intuitively, we deem that truth should be single. After all, we seem to understand Tarski's series of truth-predicates by appealing to a single truth-concept. This is why, in the end, we call all these predicates "*truth*-predicates". But the very theory at stake needs to prevent us from expressing such a concept, lest incurring in inconsistencies again.[9]

Undeniably, Tarski's revisionism is problematic. As the objections just considered show, the gulf it creates between our intuitions and a formal theory of truth is too broad. This is probably why, drawing from the critique of Tarski's proposal,

[8] "(…) The very possibility of a consistent use of the expression 'true sentence' which is in harmony with the laws of logic and the spirit of everyday language seems to be very questionable, and consequently the same doubt attaches to the possibility of constructing a correct definition of this expression."
Tarski: 1933, p. 165.

[9] The charge of fragmentation might be read as a version of the first worry: the single truth-concept here required seems to correspond to the ordinary truth-notion that underlies the semantic expressions of natural languages.

subsequent investigations of the truth-notion have tended to assume that such a gulf should be totally (as Priest contends) or almost totally (as non-dialetheists often suggest) eliminated. No analysis of truth, it is often stressed, should be utterly revisionary.[10]

In the remaining of this section, I would like to sketch a more charitable approach to Tarski's theory. The counter-intuitiveness of Tarski's solution may not require concluding that no insightful remarks about the truth-notion have been offered. It seems to me that Tarski's proposal might be read as drawing at least three interesting conclusions from the liar paradox:

(i) a global semantic predicate cannot consistently be combined with a self-referential language: universality and consistency are unable to go together, analogously to what happens in set-theory;

(ii) a consistent semantic model is somehow at odds with our naïve intuitions about truth;

(iii) once restrictions on the semantic notions of a language are introduced, an object-language/meta-language distinction is required, in order to allow a semantically richer meta-language[11].

In the literature, these points do not tend to be acknowledged as merits of a semantic theory – not all of them, at least. They often seem to be conflated with Tarski's unpalatable fragmentation of the truth-notion. However, it is not obvious that such features entail, *per se*, a radical setting aside of natural languages as incoherent; nor a hierarchy of semantically open languages. Some of the subsequent approaches to the liar paradox might in fact be read as having better preserved and clarified these three points.

2.1 Kripke's Theory of Truth

Saul Kripke has proposed a major alternative to Tarski's theory of truth.[12] Tarski's Undefinability Theorem assumed that the truth-predicate should range over all the sentences of a language. Kripke shows us that by dropping such assumption we can obtain a consistent model for a semantically closed language.

Kripke's proposal aims at defining a partial truth-predicate. It does so by means of a mathematical construction. (I shall illustrate it in a simplified way.) Roughly, we start with a basic model M for all the sentences of a language L. L is not endowed with a truth-predicate and all predicates of L are supposed to range over a fixed domain D. We next extend L to a language *L* by adding a partially defined predicate T(x), intended to be the

[10] Cf. Priest: 1987a and 1987b. For a recent non-dialetheist critique of Tarski's theory, see Scharp: ms.

[11] David De Vidi and Graham Solomon persuasively argue that Tarski does not use the expression "richer meta-language" in a clear-cut way. Cf. De Vidi and Salomon: 1999. Herein I use the expression "semantically richer meta-language" in order to refer to a meta-language that contains a truth-predicate (or related semantic expressions) which applies to sentences of the language under investigation.

[12] See Kripke: 1975. Robert Martin and Peter Woodruff offered, almost simultaneously, a similar construction. See Martin and Woodruff: 1975. I shall here confine to Kripke's version, for it has been the most influential and philosophically discussed.

truth-predicate for L. The pair of sets (S1; S2) is taken to correspond, respectively, to the extension (the whole of the sentences of D that qualify as true) and to the anti-extension (the whole of the sentences of D that qualify as false) of T(x).

The predicate T(x) is defined by means of an infinite series of stages. We define S1 and S2 by building an infinite hierarchy of interpreted languages L_n(S1*;S2*), where S1* and S2* represent, respectively, the sets of the true and false sentences for each level. At the bottom level, the set (S1* U S2*) is empty. At the next level, sentences not containing the truth-predicate are assigned to S1* and S2*. (The evaluation of compound sentences adopted by Kripke follows Kleenes strong tables.) At each further level, the extension and anti-extension of T(x) retain the sentences assigned at previous levels and are implemented with previously undefined sentences. The process continues till we reach the least minimal fixed-point, at which the set of true (false) sentences is equal to the set of true (false) sentences of the previous level.[13] Not all sentences of *L* are thereby given a truth-value.

To the sentences that remain semantically indeterminate, Kripke assigns the status of *ungrounded*. His definition is the following: given a sentence *A* of *L*, *A* is grounded iff it has a truth-value at the minimal fixed-point.[14] Sentences such as 'this sentence is true' and the Liar sentence are not assigned a truth-value at the minimal fixed-point. Hence, they are ungrounded. Furthermore, the theory allows for a mathematical definition of *paradoxical* sentences: a sentence is paradoxical iff it cannot be assigned an arbitrary truth-value at any fixed-point, on pain of contradiction. The liar sentence, among others, belongs to this category.

2.2 An Evaluation of Kripke's Theory of Truth. Trying to Handle Tarski's Ghost

Kripke's theory of truth achieves better results than Tarski's. Its main merit is that of modeling a semantically closed language without embracing inconsistencies. We have a model that gets closer to the mechanisms of natural languages: sentences can predicate their own semantic properties and one single truth-predicate is adopted. Kripke also provides us with a justification for the distribution of truth-values. Within his construction, the ungrounded sentences are those which do not refer to sentences whose truth-value depends, so to speak, on non-semantic facts. This mirrors our intuitive conception of truth-value ascriptions. If we have to assign a truth-value to sentences that involve a truth-predicate, we seem to look at the truth-value of further sentences, which in turn do not contain semantic expressions. Hence, Kripke contends, the non-assignment of a truth-value to sentences such as the Liar one is intuitively justified.

Beyond listing the benefits of his solution, however, Kripke has persuaded commentators of one major worry: the problem being that the theory does not avoid counter-intuitive results. Kripke's proposal defines the notions of 'fixed-point', 'paradoxical' and 'groundedness' for a language *L* by deploying a meta-language. This inevitable fact, however, leads to unpalatable outcomes. There are intuitively true claims about sentences of *L* that are expressible in *L*, but true only within a meta-language. For instance, given that A is the liar sentence of *L*, the sentence

[13] That a fixed point is achieved can be proven by means of Moschovakis's inductive constructions or by Zorn's Lemma. Kripke opts for the former. See Kripke: 1975, p. 66.

[14] Ibid., p. 71.

(1) A is ungrounded

cannot be true in *L* .

Another way of putting the difficulty is the following. Kripke's theory of truth falls prey to Liar-revenge problems. Kripke describes the Liar sentence as undefined. Hence, it (intuitively) seems to follow that the sentence

(SL) The Liar sentence is not true

is true in *L*. The inference, however, sends us back in paradox.[15] Kripke's proposal avoids this strengthened form of the Liar paradox by not assigning a truth-value to (SL). In *L*, the Liar sentence gets truth-value undefined. Given a correct interpretation of *L*, it follows that the sentences 'this sentence is false', 'the Liar sentence is not true', 'the Liar sentence is true' get the status of undefined at every fixed-point. Hence, Kripke adds, in order to truly assert (SL) we have to ascent to a meta-language level.

In the literature, there is widespread agreement about reading Kripke's appeal to a meta-language (or, as Kripke put it, to Tarski's ghost) as a drawback. As with Tarski's theory, commentators read the presence of more than one semantic level as an undue departure from our intuitions. An adequate model for a semantically closed language, it is assumed, has to accommodate *all* the claims that we intuitively regard as true. And it seems that the task should be achieved by means of the very truth-predicate that the theory defines. Here, I would like to challenge such an assumption. Colin Howson, in a forthcoming article, puts forward an interesting reply to the Strengthened Liar critique.[16] If his proposal is accepted, it seems to me, formal theories of truth such as Kripke's might be put in a position to handle the counter-intuitivity objection.

Howson defends Kripke's construction in two moves. First, he spells out in more detail which aspect of the proposal is hostage of the Strengthened Liar reasoning. As seen above, the reasoning infers, from the undefined status of the Liar sentence, the truth of (SL). Such an inference, however, depends on the Liar sentence belonging to the anti-extension of T(x): it can be allowed only by closing off the truth-predicate – i.e. by giving up its partiality. This shows that the Strengthened Liar reasoning is not simply directed against the incompleteness of Kripke's model. Rather, the critique stresses a distance between our intuitions and the very partiality of the truth-concept. From the Strengthened Liar, Howson argues, a lesson about the truth-concept might then be drawn. Informal reasoning about Kripke's model recalls a totally defined truth-predicate and violate partiality. As a result, we can read Kripke's model as hinging upon two *complementary* truth-notions. On the one hand, we have a partial truth-concept. As Tarski taught us, if we want a semantic theory to be consistent, partiality comes in. On the other hand, there is a global (and intuitive) truth-notion; that we adopt in order to state and appreciate the theory.

[15] Here is the paradox. Let the sentence (SL) be "The Liar sentence is not true". If the Liar sentence lacks a truth-value, then (SL) is not true (and not false). Infer by T-schema the truth of (SL). Then (SL) is both true and not true. Contradiction. Conclude that (SL) is false. By reading "false" as "not true", it follows that the Liar sentence is true – contrary to the claim that it lacks a truth-value.

[16] Howson: forthcoming.

Howson's remarks might also help us to make a further point. One might argue that the construction in question does not push unaccommodated intuitions to a meta-language level without a reason.[17] The theory tells us that some of our intuitions (the ones about the totality of the truth-notion) are cause of paradox. Also, it shows that such intuitions can be preserved, harmlessly, at a complementary semantic level. Hence, the adoption of a meta-language might be regarded as good news. It may be a useful tool for relating two complementary and now distinct truth-notions.

3.1 Barwise and Etchemendy. Truth within Situation Semantics

So far, we have seen how, on the one hand, Tarski's solution to the liar paradox helpfully points to the need of a consistent but restricted truth-predicate. Tarski's hierarchy, however, remains too much at odds with our intuitive notion of truth. Kripke's theory, I have maintained, might be read as an interesting development of Tarski's insights, hinging upon the adoption of two complementary truth-notions. In this section, I shall suggest that my reading and defense of Kripke's theory might be extended to a further semantic model, given by Jon Barwise and John Etchemendy (henceforth, B&E).[18]

B&E consider truth as a property of propositions. Their construction hinges upon three main tools. First, they appeal to a form of situation semantics: both propositions and truth-makers are modeled as sets of partial worlds.[19] Second, they make use of a non-well founded set theory, developed by Peter Aczel. Thereby, they allow self-referential claims.[20] Third, B&E account for propositions in an Austinian way:[21] each proposition is taken to say of an actual or historical situation whether it corresponds or not to a certain situation-type.

The model adopts the following primitives:

[17] I depart from Howson on this point. He concludes that Kripke's theory does not require a meta-language, but two interpretations of the same language. On the one hand, Howson contends, we need a classical interpretation, according to which the truth-predicate is totally defined. (Such an interpretation has been offered, for instance, by Feferman's axiomatization of Kripke's theory). On the other hand, we need a non-classical interpretation of the theory, according to which the truth-predicate is partially defined (and, say, Kleene's strong tables adopted). See Howson: forthcoming.

The reason of my departure is twofold. For one, if we accept to distinguish between two notions of truth I see no reason for dismissing with a meta-language any more – lest we want to pursue the implausible ambition for a universal language. For another, reasonings such as the Strengthened liar one can become inferences *about* the results of a semantic theory only if we ascent to a meta-level. (In fact, Feferman achieves a consistent classical interpretation of Kripke's theory by forbidding the inference from the lack of truth-value of the liar sentence to the negation of the non-truthood of such sentence). The Strengthened Liar inference seems to be intuitively understood precisely as a reflection about the theory under investigation. For an analogous interpretation of the Strengthened Liar reasoning, see Glanzberg: 2004.

[18] Barwise and Etchemendy: 1987.

[19] Cf. Barwise and Perry: 1983, especially parts B and C.

[20] Aczel's non well-founded set theory drops the axiom of foundation from ZFC and replaces it by an anti-foundation axiom that allows sets to be non well-founded. The theory is known to be consistent relative to ZFC. Incidentally, McLarty argues that the use of Aczel's set-theory is not essential to B&E's proposal. McLarty: 1993.

[21] 21 Actually B&E define two kinds of propositions: Russellian propositions and Austinian propositions. They explicitly regard the Austinian version as more satisfactory (Cf. Barwise and Etchemendy: 1987, pp. 97 ff. I therefore confine to the Austinian version.

i) *state of affairs*; it encodes the information about certain worldly relations, by means of an ordered (n+2)-tuple, containing any n-ary relation *R*, a series of n objects and a member *i* of the polarity (1,0), which represents the having (i=1) or not having (i=0) of the relation.[22]

ii) *situation*; it is a set of states of affairs; it corresponds to an individual truth-condition;

iii) *proposition*; it gets coded as conveying the membership of an actual situation *s* to a type *T* of state of affairs; its model corresponds to a pair {s,T} consisting of a situation and a type of states of affairs.

Within the resulting construction, truth can be defined in accordance with the following schema:

(Propositional T-schema) a proposition of the form $\{s; [\sigma]\}$ is true iff $\sigma \in s$.

For any proposition *p*, *p* is true iff there is some actual situation in which it is true that *p* is true and if the situation fits the situation-type predicated by *p*. Notably, the model thereby retains bivalence: every well-formed proposition is either true or false.[23] Moreover, the truth-predicate is given a variable extension. The domain of the variables involved in a proposition is partially determined by the situation that proposition is about. Hence, each proposition is allowed to undertake a semantic shift when it refers to a different situation.

B&E's theory allows for Liar-type propositions. For every situation *s* there can be a Liar proposition *fs*, about *s*. The proposition gets modeled as

$$fs = \{s; [Tr, fs; 0]\},$$

where 'Tr' is the property of being true and *fs* is interpreted as saying of a situation *s* that this very proposition does not hold in *s*.[24] In the modelled language, the proposition *fs* is an element of situations such as the situation *s* ∪{falsity of *f*}). However, *fs* is not an element of the situation *s* it is about. Nor, B&E contend, there can be a situation encompassing all facts: in fact no situation is about the whole world. The result of these measures is that each liar proposition "diagonalizes out of the situation" it is about.[25] We can conclude that any proposition of type *fs* is false: no paradox.

[22] Barwise and Etchemendy: 1987, p. 30.

[23] Inasmuch as truth is taken to be a property of propositions, it might be noticed, bivalence is hard to abandon. Standardly, the meaning of a proposition depends on its having a truth-value. Either a proposition is true, or it is false. If a sentence fails to express a proposition that can be regarded as true or false, it fails to convey a proposition *tout court*.

[24] B&E also define, for each situation *s*, a Denial Liar *d* that denies its own truth – where *d* asserts its own falsity. See Barwise and Etchemendy: 1987, pp. 16 ff., and 122. In the main text I confine to what they call "the assertive liar" both for space reasons and for that should suffice for the purpose of my discussion.

[25] Cf. Barwise and Etchemendy: 1987, p. 154.

3.2 An Evaluation of B&'s Model. A Benign Semantic Ascent?

The semantic theory offered by B&E has at least two original features. For one, we have a semantically closed language whose model does not abandon bivalence. For another, the expression 'is true' is allowed to undergo semantic shifts, thereby reflecting our interpretative processes. It is intuitively true that a shift in meaning occurs between the assertion of, say, the Liar proposition and the appraisal of its truth-value. B&E's model grasps this feature: it allows any Liar proposition to be false for the very situation *s* that proposition is about, and true for a subsequent situation *s'*.

Its achievements notwithstanding, also B&E's model has been charged of counter-intuitivity. The difficulty might be spelt out in two ways. First, it seems difficult to accept a situation-dependent notion of truth. Truth, intuitively, is not relative. A confirmation being the fact that the very (meta-)language in which the theory is stated hinges upon a total truth-notion.[26] A second objection has it that B&E's construction suffers of liar-revenge problems. The model retains consistency at the cost of not accommodating all intuitively expressible propositions. As briefly mentioned above, B&E's diagnosis of the liar paradox needs to banish Liar propositions that talk about the whole of all facts. Otherwise, contradictions would follow. But why do we need to rule out propositions about the whole world? No intuition tells us that such propositions should be banished.

B&E do not address both objections. They dismiss the first one by explicitly confining their task to an analysis of the logical aspects of a language. They make it clear that they have no intention to give a complete account of the mechanisms of a language; nor to take into consideration the relation between their model and, so to speak, "real propositions".[27] The move, no doubt, is legitimate. However, it leaves our worry intact. To the second objection, B&E offer a more extensive response. They contend that the limitation of expressive power entailed by their theory is less problematic than it *prima facie* appears. They prove a theorem, the Reflection Theorem devised to show that Austinian propositions can be about very broad situations, so to encompass all the Russellian facts of a world.[28] The answer, nevertheless, might be regarded as unsatisfactory. For it does not really address the worry from which the objection was raised. The reason we are unhappy with a restriction of expressive power is that we want propositions to be possibly global. No matter how concessive the restriction is, intuitively acceptable propositions will be probably ruled out.

It seems to me that B&E's model might better handle both counter-intuitivity worries by adopting a different stance. As suggested for Kripke's theory, the model might be taken to show that we need two complementary truth-notions. On the one hand, B&E define a context-dependent truth-predicate. The measure blocks the liability to diagonalization of our (naïve) truth-concept. B&E steer clear of inconsistencies by relativizing truth to partial (model of the) worlds, analogously to Kripke's introduction of a partially defined truth-predicate. On the other hand, B&E's theory does not make commitment to a radical dismissal of a total truth-notion. In order to state their theory, they make claims about *all* the semantic facts of the modelled language. Moreover, insofar as we intuitively understand truth as a global, this is the truth-concept that we

[26] Cf. Priest: 1993.

[27] Barwise and Etchemendy: 1987, pp.189 ff.

[28] Ibid., p.157.

need to use in order to grasp and assess the theory. An ascent to a meta-language, if this reading is accepted, might then be allowed as a benign device, voted to lodge the latter truth-notion.[29]

Conclusion

I have argued that the counter-intuitivity charges to formal theories of truth might be less harmful than it is usually assumed. Tarski's appeal to a manifold of truth-predicates is unpalatable. Nevertheless, his analysis of the truth-concept contains interesting insights, which subsequent models seem to clarify. Analogously to what Tarski did, Kripke's and B&E's semantic models (i) put forward a restricted truth-predicate, (ii) take distance from our naïve conception of truth and (iii) adopt a meta-language that hinges upon a broader truth-notion. Differently from Tarski's solution, however, those two models can be provided with an explanation for their departure from intuitions. They can be read as propounding a doctrine of two complementary truth-notions. On the one hand, both models appeal to a partial truth-concept in order to achieve consistency. They also deploy, on the other hand, an intuitive and global truth-notion: this latter notion being required in order to state and assess the theories. Interestingly, the appeal to a complementary linkage between an intuitive and a (partially) revisionary truth-concept may be taken to represent the very lesson of the liar paradox. The diagnosis might be that two notions of truth need to be adopted. And carefully kept apart.

There is no intention to claim that the analysis of the positions here considered has been exhaustive; or to maintain that the reasons given in favors of Kripke and B&E's models may be conclusive. Rather, I hope to have given support to the thesis that a complementary relation between formal theories of truth and our intuitions might represent an interesting route for approaching the liar paradox.

Simona Aimar
St. Andrews University

[29] Priest-wise objection: what about truth within the meta-language? If we allow for a total truth-predicate, we are back in paradox. Reply 1: I see no need for the meta-language to be semantically closed, inasmuch as I take it as an interpretative tool for the theory. Reply 2: one might also allow for a semantically closed meta-language, given that this is axiomatized so not to allow propositions about its whole domain. Inasmuch as the meta-language is a language devised *for* the object language, the measure should not be regarded as unsatisfactory.

References

Barwise, J. and Perry, J. Situations and Attitudes (Cambridge: Cambridge University Press, 1983).

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

DeVidi, D and Solomon, G. "Tarski on 'essentially richer' meta-languages," Journal of Philosophical Logic 28 (1999): 1-28.

Howson, C. "Truth and the Liar," (forthcoming).

Kirkham, R. Theories of Truth (Cambridge, Mass: the MIT Press, 1992): ch. 9.

Kripke, S. "Outline of a Theory of Truth," Journal of Philosophy 72 (1975): 690-716. Reprinted in Martin, R. Recent Essays on Truth and the Liar Paradox, (Oxford: Oxford University Press, 1984): 53-83.

Glanzberg, M. "Truth, Reflection, and Hierarchies," Synthese 142 (2004): 289-315.

Martin, D. 'Review of The Liar. An Essay on Truth and Circularity', The Journal of Symbolic Logic 57 (1992): 252-254.

Martin, R. Recent Essays on Truth and the Liar Paradox, (Oxford: Oxford University Press, 1984).

Martin, R. and Woodruff, W. "On representing 'true-in-L' in L," Philosophia 5 (1975): 213-217.

McLarty, C. "Anti-foundation and self-reference," Journal of Philosophical Logic 22, (1993): 19-28.

Priest, G. In Contradiction, (Dordrecht: Kluwer Academic Publishers, 1987): ch. 1.

Priest, G. Unstable solutions to the Liar Paradox, (Self-Reference: Reflections on Reflexivity, Dordtrecht, Boston and Lancaster: Martinus Nijhoff Publishers, 1987).

Priest, G. "Another disguise of the same fundamental problems: Barwise and Etchemendy on the Liar," Australasian Journal of Philosophy 71 (1993): 60-69.

Scharp, K. "Fragmentary theories of truth," (manuscript).

Tarski, A. "The Concept of Truth in Formalized Languages," Logic, Semantics, Meta-mathematics. Trans., J. H. Woodger. Oxford: Clarendon Press (1956): 152-278.

Tarski, A. The Semantic Concept of Truth Readings in Philosophical Analysis. Eds., Herbert Feigl and Wilfrid Sellars (New York: Appleton-Century-Crofts, Inc., 1949): 52-84.

Causal Relevance in Brain Simulators

Eve Rips
Stanford University

ABSTRACT

In this essay, I argue that artificial intelligence can be broken into two subcategories: simulator AI and functional AI. To count as an instance of simulator AI, the system in question must possess intentional states through the same causal process that a human mind does, while functional AI requires a different causal mechanism than humans possess. This distinction is based on arguments from Jaegwon Kim and Lawrence Schapiro which claim that multiple realizations are only possible when a new causal mechanism is involved. Consequently, it can be possible even for theorists like John Searle who reject functionalism to accept the possibility of simulator AI.

In response to objections against the Chinese Room argument, John Searle argues that brain simulators made of water pipes, or perhaps even silicon, would not be able to possess consciousness. I argue that it is possible for even a type identity theorist to reject this idea. Identity theorists allow for some differences between brains -- properties that are not causally relevant to the scientific kind involved have room for variation. In the case of being able to have intentional states, this seems to allow for differences in the size and color of involved nerves, and leads to the conclusion that some variation in the elements involved in these processes could be possible. Philosophers like Jaegwon Kim and Lawrence Schapiro have argued that for a system to count as a new realization of a mind, it would have to be a different scientific kind, and consequently that the system would have to operate in a causally different way than human minds. Ultimately, AI should be thought of as having two sub-categories – simulator AI and functional AI. Simulator AI would consist of any system that worked through the same causal mechanisms as human minds, while functional AI would require that a system was genuinely a new realization of a mind, and therefore worked through a different causal process. The former can be defended without attacking identity theory, but the later cannot.

I make this argument in five stages. In part one I describe Searle's famous Chinese room argument. In the second part, I clarify two aspects of Searle's argument – that it is fundamentally an attack on functionalism, and that it makes claims about *intentionality,* and not just understanding. Next I describe the causal powers argument advanced by Jaegwon Kim and supported by Lawrence Schapiro. I then attempt to show that carbon is not causally relevant towards understanding, and consequently that a brain made out of silicon would not constitute a new realization of a brain. Lastly, I breakdown AI into simulator AI and causal AI, and argue that one could embrace the possibility of simulator AI while still rejecting functionalist theories.

The Chinese Room Argument

At the heart of the debate on strong artificial intelligence is the question of the causal mechanism behind phenomenal consciousness. Such was the case for John Searle in his seminal Chinese room argument, and the defenses he subsequently advanced in its favor. In these arguments, first proposed in his article "Minds, Brains, and Programs," Searle argues that strong AI is impossible, even if one were to build a neuron for neuron model of a brain out of silicon instead of carbon.

Searle's focus is on understanding Chinese as a representative example of strong AI. The argument is posed as an attack on the machine-state functionalism advocated by Alan Turing, and other believers in the Turing Test theory that a system that was input-output equivalent to human minds would be capable of understanding. Searle imagines himself locked in a room with a complete set of rules for Chinese inputs and outputs. If a Chinese speaker inputted a card with specific Chinese characters, Searle would be able to go the book, find the input, and read the rule telling him which card to output. If the rulebook were accurate enough, the Chinese room would be input-output equivalent to speaking with a human who was fluent in

Chinese. Despite this, however, Searle would have no understanding of the Chinese language himself. The job could be completed without any knowledge of what the specific symbols in question meant. Thus the system would be unable to be distinguished from a real Chinese speaker, but there would, nonetheless, be no real understanding involved.

As a rejection of the Turing arguments, Searle's article is remarkably successful. A machine such as the one described by Searle would be able to fool Turing (assuming Turing could speak Chinese). Indeed, the claim advanced by Turing that any machine able to fool us into believing it can think is genuinely able to think seems problematic in several ways. Who is the "us" the machine has to fool? Is it just Turing himself? Is it anyone of comparable intelligence and gullibility to Turing? Is it all members of humanity? If we were to discover a tribe of people who were particularly easy to convince of things and were fooled into believing light switches could think, would that mean a light switch had understanding? Clearly the Turing test faces serious difficulties on several counts, and Searle seems quite right to reject it. As a rejection of the possibility of strong AI and machine-state functionalism, however, Searle's arguments are less successful. Searle is not arguing simply that Turing equivalency cannot predict when machines can think, but instead that systems like the Chinese room could never think.

In response to his initial publication of the Chinese Room argument, several objections were advanced. The Berkeley and MIT objection argues that, though there might not be thinking in the case of the Chinese room, a brain simulator could be created that contained the same patterns of neural firings that one finds in a human brain. This simulator, though it would not actually be made of the same material as a human brain, would nevertheless be able to think. "At the level of the synapses," the reply, as Searle explains it, demands, "what would or could be different about the program of the computer and the program of the Chinese brain?"(514).

Searle counters this reply by imagining a different room in which there is a man who, upon receiving an input, looks up in a rule book which valves to turn on and off in a series of water pipes set up to mimic the neural firings that would occur for a native Chinese speaker. He then follows this rule and gets the right output from the water pipe machine. "Where," Searle wonders, "is the understanding in this system?"(514). Searle's implication that inability to locate understanding within a system makes it impossible for there to be any understanding is deeply flawed. The question of where understanding is located is a puzzling one – the explanatory gap has been a problem for philosophers for centuries. Even identity theorists, though convinced that understanding is reducible to a complex series of neural firings, are currently unaware of exactly which particular neural firings are essential to understanding Chinese. Searle's indignant "where's the understanding in that?" could be just as easily demanded of human brains. If Searle's implication that inability to locate understanding in a system makes it impossible for the system to understand things were correct, then we would be a race of zombies! Fortunately this line of reasoning does not seem acceptable.

Searle attempts to delve deeper into the "where's the understanding?" question by analyzing the individual components of the water pipe system:

> the man certainly doesn't understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the conjunction of man and water pipes understands, remember that in principle the man can internalize the formal structure of the water pipes and do all the "neuron firings" in his imagination. (514)

For Searle, the distinction between the Anglophone-rulebook-water-pipes situation and the human brain situation seems to be in the stipulation that the pipes situation simulates only the structure of the nerve endings. "As long as it simulates only the formal structure of the sequence of neuron firings at the synapses," he argues against neuron simulation systems, "it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states."(515). Searle feels that simulations of the brain that go beyond the formal neural structure cross the line of understanding "solely in virtue of being a computer with the right sort of program."(517). That is, if a program too closely simulates a human brain, it can no longer constitute an instance of artificial intelligence.

This distinction is problematic, in part because Searle does not make the concept of "formal structure" sufficiently clear. When Searle stresses that the brain simulator cannot go beyond the "formal structure of the sequence of neuron firings at the synapses,"(515) he explains neither why a brain simulator can only embody formal structure nor what constitutes this structure. So long as his claim that "if you can exactly duplicate the causes, you could duplicate the effects,"(517) stands, then what is to prevent the creation of a system of rule-operated pipes being created that is an exact duplication of the causes within the mind of a native Chinese speaker? Admittedly, this pipe system is a more absurd idea than, for instance, a silicon replication of the brain, but it is a theoretic possibility. Unless Searle is arguing that thought can only come from the particular combination of the seven chemical elements that make up the human brain, then "exact duplication" of Chinese thought processes with an Anglophone and a series of water pipes might be possible.

Searle seems to hold that a system would only be able to be classified as strong AI if it is somehow causally different from a human brain. He allows that an exact man-made replication of a brain would be able to think. Searle seems to think that brain replications made of water pipes or even silicon would be causally different from brains, and therefore would not constitute real intelligence or thought. He seems to believe that somewhere in this shift to new building materials understanding would be lost. This faces two problems – it fails to account for why silicon or water pipes would be causally different from carbon and it fails to explain why, even if they were causally different, they don't possess mentality. I argue that Searle is wrong to assume that there are essential causal differences between systems made of carbon and systems made of silicon or water pipes.

Two Clarifications of Searle

Before attempting to address this problem, two important clarifications of Searle's arguments need to be made. First off, though Searle frames his argument as a commentary on artificial intelligence, his argument is fundamentally an attack on functionalism and a defense of a physicalist perspective. Searle only explicitly mentions functionalism twice in the entirety of "Minds, Brains, and Programs." In both cases he makes a statement about AI and includes functionalism as a parenthetical, such as in the statement "In strong AI (and in functionalism, as well) what matters are programs" (521). Nevertheless, the Chinese room argument serves as an attack on the general functionalist perspective, and not simply on strong AI.

The other important clarification of Searle's argument is that he is really making claims about intentionality, and not just understanding. It has always been difficult to decide which aspects of mentality to focus on when discussing the possibility of AI and, indeed, many other problems in the philosophy of mind. Ability to feel pain and ability to understand Chinese may both be indicators, or perhaps even conclusive proof, of mentality, but nevertheless the two examples are quite different. Understanding Chinese seems like a good property to focus on when discussing strong AI, since it involves both knowing complex rules and somehow being able to comprehend them. However, *understanding* is something of a confusing term. One could interpret 'understanding' as "perceiving meaning in," "having a semantic interpretation of," or simply "being familiar with." What Searle seems to be addressing is really the property of intentionality – having mental states which are about things. Though I continue to use the term 'understanding' to describe the property at the heart of the Chinese room argument, it should be kept in mind that the term is meant to refer to possessing intentional states on the subject at hand. With these two clarifications in mind, we are better armed to discuss the role causal structures play in discussion of AI.

The Causal Powers Argument

There are two ways in which one could argue that a brain made of water pipes or silicon could be capable of thinking – one could argue for a functionalist conception of multiple realizablity and then defend the idea that a silicon or water pipe brain could fulfill specific functional roles, or one could argue that a brain simulator could, theoretically, not count as a novel realization of human brains as we currently know them, and that consequently one would not have to embrace functionalist theories in order to believe in artificial intelligence. Attempting the functionalist defense could beg the question, and is therefore not an acceptable option -- Searle is, after all, making an argument that rejects functionalism, so a counterargument that assumes multiple realizability is possible would be problematic. This second idea, that a silicon brain might not count as a new realization can successfully be defended by Jaegwon Kim's causal powers argument and Lawrence Shapiro's corkscrew argument.

To argue for the functionalist conception would be rather obviously question-begging. The question at the heart of the Chinese room argument is whether a machine that was functionally indistinguishable from a native Chinese speaker would also possess the understanding that the native speaker possesses. As argued before, Searle's argument is distinctly anti-functionalist. To respond to an argument rejecting machine-state functionalism that it fails because it does not allow for multiple realizability would be absurd to say the least. To uphold this line of argument one would have to provide a more elaborate defense of functionalism as a whole. Better, then, to attempt to bypass the difficulties inherent in assuming multiple realizability, and argue instead that one can have certain types of variation among brains without necessarily having a new realization.

In his article "Multiple Realization and the Metaphysics of Reduction," Jaegwon Kim advances what is now known as the Causal Powers Argument. Kim argues that scientific kinds are divided by their causal powers. Kim believes that one can use this principle to argue that different physical realizations of a mental kind can be thought of as different kinds. Accordingly, mental kinds for which structure is not affected are not different causal kinds, and so cannot be considered to be different scientific kinds at all, since scientific kinds are divided by causal powers. Ultimately, then, Kim holds that differences in mental kinds that are not causally relevant do not count as distinct mental realizations, and that differences in mental kinds that *are* causally relevant do not count as multiple realizations of the same kind, because causal differences cannot exist within a scientific kind. Thus, much as jadeite and nephrite are not the same scientific kind, a human and a machine that operate in physically different ways would not be the same kind of thing. If the machine was causally the same as the human then it wouldn't be a novel realization of mentality, and if the machine was causally different, then it wouldn't be in the same scientific kind as a human mind.

This argument is similar to one made by Lawrence Shapiro in his article "Multiple Realizations." Shapiro argues that, when discussing specific functional kinds, there can only genuinely be multiple realizability when changes to causal structure are involved. Two corkscrews that were different colors would not count as multiple realizations, because the causal process that allowed them to open bottles of wine would be identical. The same would be true if one corkscrew was made of steel and the other of aluminum. However if one corkscrew worked by leverage and one worked by dropping small quantities of potent acid on the cork, the two devices would count as different realizations, because different causal mechanisms would be involved. Similarly, human brains that were slightly different sizes would not be different realizations of mentality, but a human brain and a computer program perhaps might be.

Admittedly both of these arguments have problematic aspects. Schapiro in particular seems to assume a degree of functionalism – he focuses on the role of corkscrews as a functional kind, and suggests that this can be applied to discussion of the mental. Identity theorists would argue that it makes no sense to talk about these sorts of functional roles in the context of mental states. While corkscrews are a functional category, defined by their ability to remove a cork from a wine bottle, mental properties like pain are defined not by their functions but by being c-fibers firing. However, identity theorists do still have to account for the problem of how two people with very different c-fibers can both have pain, and the idea of scientific kinds is a natural way to address the problem.

A devoted identity theorist could still use causal roles in discussing mentality – even if she argued that multiple realizability was completely impossible. Even identity theorists allow for some variation in what it is to be a pain – the c-fibers in my brain are not the same as the c-fibers in my friend's brain or in the brain of my dog. The physicalist has to find a way of grouping all these states together. This can be done through arguing that my c-fibers and my friend's c-fibers are the same scientific kind. Because scientific kinds are determined by causal properties, the physicalist can argue that the two sets of c-fibers are not multiple realizations of pain. When faced with the question of whether a machine could ever be in pain or possess understanding, a physicalist could attempt to argue that no causal similarities are possessed, and therefore that it makes no sense to talk about the pain or understanding of the machine, since it pain and understanding are properties of c-fibers.

These arguments seem to get at the heart of Searle's arguments about formal neural structure. When Searle talks about simulations that go beyond formal neural structure, and argues that they don't count as understanding "solely in virtue of being a computer with the right sort of program" (517), what he actually seems to be saying is that a structure which was causally identical to human brains would not be a new realization of the brain, and would not genuinely count as instances of artificial intelligence. That is, to count as artificial intelligence, one cannot simply create an exact replica of a human brain. There has to be something genuinely *artificial* about the system in question. Searle seems to believe that the actual physical matter that composes minds is causally relevant. Something about the mix of carbon and the six other biological elements that make up human brains is essential for there to be understanding.

Is Carbon Causally Relevant to Understanding?

The causal relevance of carbon to understanding is by no means as apparent as Searle makes it seem. A functionalist might ask how it could be conceivable to know that carbon is essential to understanding. What is there to prevent a Martian from having understanding despite having a mind made of the element greenslimium? Searle seems to rely on his knowledge of mental states as we know them to answer that question. All currently known instances of mental states involve carbon, so he presumes carbon to be essential for mentality.

Knowing what causes mentality may be impossible to empirically ascertain. While a functionalist would say that an input-output equivalent brain made of silicon instead of carbon possessed understanding, a type identity theorist would argue that what it is to be a have mentality is defined in terms of being made of carbon, and that therefore the silicon brain has no mentality. Proving what causes mentality is therefore only partially an empirical question.

But let's examine the question from the physicalist perspective -- thoughts are c-fibers firing. If the c-fibers in question are slightly larger or a darker shade of gray than is usual for brains, this doesn't seem to cause a particular threat to the mentality of the individual in question. After all, there still are c-fibers, and they are still able to fire. The question of "what defines the category of things that have pain" can be reduced to "what defines what it is to be a c-fiber." When physicalists allow for different lengths of c-fibers to be included in the same scientific cause, there is an acknowledgement that the different c-fibers share causal powers. If physicalists allow for c-fibers to be different sizes and different colors, why could they not also be different elements?

Carbon and silicon are in the same group (column) on the periodic table – the carbon family. This means that while they possess different numbers of protons, and have different chemical weights, they both have four electrons in their outermost electron shell, which is a p2 orbital. This means that, as with all periodic table groups, the properties the elements will have in chemical reactions are remarkably similar. Just as most things, including c-fibers, could have their color changed without affecting any causal relations, so too could most things have a particular element switched to another element in the same period without affecting any other properties besides weight. Thus it seems very reasonable to believe that a replication of a human brain could be created out of silicon instead of carbon without any important changes. Physicalists allow carbon c-fibers to be heavier in some brains than others because some possess more carbon atoms than others – it seems reasonable, then, to allow them to be heavier because the elements involved contain more protons. Imagine two brains with roughly the same arrangement of neurons. One, made of carbon, belongs to a normal human, but happens to weigh substantially less than an average human brain. The other, made of silicon weighs the same amount as a normal brain. Why side with the importance of proton number over total weight as a cause of mentality?

Of course, this is not conclusive proof that a silicon brain could ever possess understanding. But conclusive proof of understanding in all but the case of our own minds is exceptionally difficult. If one grants the degree of inference to the best explanation often used to justify the existence of other minds, though, perhaps one could also use such inference to allow for mentality in a silicon replica of a human brain that was roughly input-output equivalent to a human brain. Yes, there is chance that the atomic weight of carbon is essential to the consciousness of a human brain, but it does not seem like a significant chance.

A brain made out of water pipes instead of silicon would, naturally, be a more complicated issue. Whereas carbon and silicon are fundamentally very similar, bonded carbon atoms and water pipes are quite different. Thus, the question of whether a water pipe simulator of a brain would possess the same understanding as a carbon brain is a more difficult question. However, it seems odd to insist that understanding in such a system would be impossible. How could a physicalist ever prove that if something similar to a carbon c-fiber could be made out of water pipes it would not count as a c-fiber? Ultimately, then, there is a strong case for mentality in some brain simulators, and no strong case for rejecting the possibility of mentality systems that more radically different from human brains.

What Does this Mean for the AI Debate?

I suggest a distinction between two types of artificial intelligence – simulator AI and functional AI. Simulator AI would attempt to build a structure out of non-organic matter that functioned as an exact causal replica of a human brain and was capable of understanding, while functional AI would require a more dramatic change in the lower-level causal mechanisms involved.

Is it even reasonable to refer to simulator AI as a genuine case of artificial intelligence? Searle thinks not. In response to the brain simulator reply, he points out that

> it is an odd reply for any partisan of artificial intelligence (or functionalism, etc.) to make: I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works. The basic hypothesis, or so I had supposed, was that there is a level of mental operations consisting of computational processes over formal elements that constitute the essence of the mental and can be realized in all sorts of different brain processes, in the same way that any computer program can be realized in different computer hardwares. (114)

Of course, Searle goes on to deny that even if a brain simulator counts as strong artificial intelligence, it is not sufficient to produce understanding, and he advances the water pipe argument with which we are now familiar.

Searle's point seems like a reasonable critique of using brain simulators as an argument for functionalism, but does not necessarily allow for a broad enough definition of artificial intelligence. Searle's argument is an argument against machine-state functionalism and is only indirectly an argument against artificial intelligence. As a result, he may be thinking of artificial intelligence as requiring the sort of multiple realizations inherent in functionalist pictures. However, the conception of artificial intelligence does not necessarily require causal differences.

Instead, strong artificial intelligence should be defined around three properties – being created by humans instead of the traditional biological processes, being made of non-organic material, and being able to have intentional states. The first two aspects of this definition ensure that the system is artificial, and the last aspect means that the system is in some sense intelligent. Thus, a successful silicon or water pipe replication of a human brain would possess artificial intelligence, and the concept of simulator AI is justified.

Whether causal AI is possible is a more complicated question, and is beyond the scope of this paper. Although I have defended the possibility of simulator AI without taking a stance on functionalism or identity theory, it does not seem possible to be as neutral when arguing for causal AI. Functionalists could argue in defense of causal AI on the grounds that such a system could be input-output equivalent to human brains, while (type) identity theorists could make a case that intentionality can be understood as a scientific kind, in which systems must share causal powers to be viewed as a part of the kind.

Eve Rips
Stanford University

References

Kim, Jaegwon. “Multiple Realization and the Metaphysics of Reduction,” Philosophy and Phenomenological Research 52 (1992): 1-26.

Searle, J. "Minds, brains, and programs" The Nature of Mind, Ed. David Rosenthal. (Oxford: Oxford University Press, 1991).

Shapiro, Lawrence. “Multiple Realizations,” Journal of Philosophy 97 (2000): 635-654.

A Modest Argument for the Hedonistic Account of Pain

Joshua M. Mitchell
University of Virginia

ABSTRACT

Though there have been many criticisms pertaining to Epicurean ethics, I elaborate on a particularly interesting connection that Cicero illuminates concerning the absence of pain and pleasure. In section II, I outline the Ciceronian account of the argument (that the absence of pain is the greatest good), and point out its shortcomings. In section III, I offer my own reformulation of the argument from a slightly altered perspective of what constitutes pain. In sections IV and VI, I deal with various objections. Section V revisits the (once) problematic *two lives* objection, where we see that the problem ceases to exist.

I. Introduction

As we see from Cicero's account of Epicureanism, its ethical system (i.e. Hedonism) revolves around the entities of pleasure and pain. As in all ethical theories, there is a "greatest good" that is the aim of life[1]. For the hedonist, this is the absence of all pain, and they hold that this is the highest pleasure (and pleasure is equated with good for the Epicurean). There are many reasons the Epicureans give (via Cicero's testimony) for this, which involve our senses and instinctual responses to good and bad, from our moment of conscious experience (57). Also, we see via Cicero a demonstrative argument for this claim after the more informal reasons are given (59). However, this idea, in virtue of itself, has been one of much contention (which we see through schools such as the Stoics to even modern day philosophers such as Joel Feinberg). Thus, a renewed philosophical inquiry into this matter is needed.

In this paper, I shall argue three things. First, I shall show that the argument that the absence of pain is the highest pleasure, as stated by the Epicureans (via Cicero), is somewhat problematic due to a lack of adequate justification for the second conclusion. Secondly, I shall restate the argument in such a way that avoids the problems from which the former suffers (and subsequently I shall deal with certain objections that may arise from the restatement). Thirdly, I shall show that given the argument, the intuitive notion, that the *two different lives* are not equivalent, is quite misguided[2].

II. The Argument Presented by Cicero—Mind the Gap

Cicero describes the Epicurean argument that the absence of pain is the highest pleasure as follows:

(i) "When we are freed from pain we rejoice in this very liberation from and annoyance…"

(ii) "…Everything we rejoice in is a pleasure."

(iii) ∴ "It is right to call the absence of pain pleasure[3]."

(iv) Furthermore, the absence of pain is the highest pleasure.

We see that the inference from (i) and (ii) to (iii) seems valid[4]. However, this does not seem to be the case concerning the inference to (iv). Thus, this is a rather serious issue for the argument as described by Cicero, if this is how we are to take their argument

[1] This should not be confused with the quality of prudence. While the Epicureans do in fact claim that prudence is the highest good, it is obvious that in that sense, "good" is taken to mean "quality," and here "good" is meant as something obtainable, as well as being a goal. Also, needless to say, I am discounting any moral theories that champion complete and utter moral apathy, but rather the ethical theories forwarded by Aristotle, Pluto, the Skeptics, and the Stoics.

[2] For further explanation on this, see page two.

[3] Taken from p. 59 of Hellenistic Philosophy (see references)

[4] We find this to be a form Iaaa syllogism, where F = all instances that are free from pain, R = instances where we rejoice, and P = instances that pleasures. The syllogism is thus: F a R, R a P, ∴ F a P.

for the absence of pain being the highest pleasure. From the argument, it rightly follows that the absence of pain is *a pleasure* (for any *X*, so long as it is something we rejoice in, would rightly be classified as *a pleasure*). Yet, there is nothing substantial here that demonstrably shows that (iv) follows from (iii). Indeed, Wilbur may rejoice in the fact that his in-laws left his house after a month-long visit. Obviously (and quite understandably), this would be a pleasure. However, one would hopefully have reservations to claim, furthermore, that the departure of one's in-laws is the greatest pleasure.

The fourth premise is simply not given adequate defense by Cicero (indeed, we see no explicit justification as to why Cicero would think this where he writes the Epicurean argument). The most vivid way this inadequacy has been illuminated is by the *two lives* thought experiment. Namely, imagine two agents who have identical lives (and are fortunate enough to have no pain whatsoever in their lives), sans for one thing. In life one, we see that the agent has the absence of pain, but only eats bread and drinks water. In life two, we observe the same absence of pain, but the agent eats filet mignon and drinks wine. Intuitively (especially since the Epicurean wants to hold some sort of pleasure calculus in his ethics (58-59), one will want to hold that the second life is more pleasurable than the former. Doing so will then necessitate that the absence of pain is *not* the highest pleasure possible. The power of intuition wielded by this example is of a considerable magnitude. Indeed, it leaves one pondering what the Epicurean has with which to retort.

To avoid this objection of intuitive inequality (and consequently avoiding the accusation that the absence of pain is not the highest pleasure), the Epicureans introduce different kinds of desires (namely, natural/necessary (e.g. food in general), natural/unnecessary (e.g. filet mignon), and unnatural/unnecessary (e.g. a statue being erected in one's honor)), and consequently state that these different types are ranked in order of importance. Because filet mignon is natural, *but* unnecessary, it should not be so highly valued and accommodated for. Thus, they believe the desire for bread and water, and the desire for steak would be different—at least to some degree. However, this should seem questionable to us. Either way, simply because one of the entities is expendable (even when *suggested* that such an action of expending be taken) does not mean that it negates the value of pleasure that is gained from such an entity. Thus, this does not seem to be an adequate response for the Epicurean.

III. A Modest Reformulation from an Alternative Perspective

However, there are alternative ways of deriving what the Epicurean ultimately wants to hold without running into the issues of intuition that have been discussed. I claim that a) it is not necessary to assert this hierarchy of different desires as some "corollary" along with the very succinct argument given by Cicero, and b) all that is needed is a simple restatement of the argument, which focuses on speaking about pain in terms of unmet desires. Observe:

(i) If one has pain, then one has a desire that is not met. [P→D][5]

[5] This should not be confused with a causal conditional. One should take it, rather, as a logical connection.

(ii) Via modus tollens, ~D→~P.

(iii) If there are no unmet desires, then all desires of the agent have been met at time *t*.

(iv) Since ~P follows from ~D, we know that all desires of the agent have been met at time *t* when ~P.

(v) Pleasure ensues when desires are met.

(vi) ∴ ~P is where all possible pleasure has been obtained by the agent at time *t*[6].

(vii) If there are no other possible pleasures obtainable for a given entity, then that entity should be thought to be experiencing the highest possible pleasure.

(viii) ∴ ~P will the highest possible pleasure for any agent *X*.

Obviously, the success of the argument hinges on whether or not *pain* was defined in a correct manner. Yet, we are able to see a definite correlation between the presence of pain and unmet desire, which gives way to the conclusion that the unmet desire is a considerable factor in the presence of *pain*. For example, when we see one suffer pain due to unrequited love, it is not simply because the "lady fair" has rejected the poor pursuer, but is due to his desiring her in the first place (for, if he had not desired her at all, there would be no pain—and consequently no poetry by John Donne). When one suffers pain due to being poor, it is not because of the *fact* that he is poor, but because he has the unmet desire *not* to be poor, and subsequently the desires to obtain things such as steak, wine, good clothing, etc. that cannot be met when one suffers from a lack of financial wellbeing. Note that if it was because of the fact that he was poor *alone*, then he would cease feeling pain if those around him considered him to be wealthy, regardless of whether or not he had the ability to spend his money. However, most would agree that this would not occur. Further evidence is seen when observing the ideal Stoic. The ideal Stoic does not experience pain because he has accepted whatever occurs in nature as necessary. Since his *desire* is to live in accordance with nature (and the nature of the universe is not something that is significantly altered (nor its "plans" thwarted) as we can tell), pain will not be known to him. However, if the ideal Stoic desired to live in any other fashion, it is quite plausible (even likely) that at some point in time he would experience pain.

Once one has looked further into the instances where pain is present, it becomes apparent that this absence of met desires has a significant correlation with observed pain. Given the logical connection we see in the first premise (and hopefully justified by the examples above), having the situation of *unmet desires* be a falsity will necessitate the absence of pain. So long as one has accepted (i), (ii) should not require further

[6] Note that "possible pleasure" does not mean *logically possible*. Here, possible means all desires have been met, and since this is so, then it would not be possible to gain any more pleasure at *t* given (v). Of course, one may object that this assumes that the only way one is able to obtain pleasure is to fulfill one's desires. This objection is dealt with on page eight.

elaboration. Although it seems obvious, it is worth explicitly stating that since ~P follows from ~D, (iv) following from (iii) should not be problematic for anyone. The rest, consequently, *should* follow without serious issue (and if there are, then the issues will arise out of objection to Epicurean ethics, and not to the argument itself).

IV. Objections

However, there will inevitably be at least one objection that arise from this argument. One may argue that (i) is problematic in the following manner: *It seems quite plausible that there are certain pains which are produced from other sources than unmet desires—namely physical pains*. This would of course discredit the first premise of the restated argument. This objection certainly has empirical weight to it (e.g. biological and neurological studies), and thus should be dealt with accordingly.

From a physiological standpoint, it is quite difficult to deny the existence of *physical pain*. Indeed, the firing of certain nerves, such as C-Fibers, that send a response to the brain is a concrete example of what *pain* is. This, of course, is something that will occur in situations beyond any agent's control. However, with every physical pain, along with it comes the desire for that respective pain sensation to go away[7]. When Dilbert touches a hot iron, he cannot help the physiological phenomena that occur within him (e.g. the firing of C-fibers, his yelping, etc.), nor the interpretation of these phenomena as *pain*. However, almost immediately, we can imagine (and rightly so) that he then has the desire for this sensation to go away. So long as this desire is unmet, he will continue to have pain. Yet, as soon as the desire is quelled, the pain is no longer present (for if the person was still experiencing pain unwillingly, he would undoubtedly still desire for it to be removed). Therefore, it seems that the physical pain can exist and (since it brings with it a desire for its removal) still be compatible with the first premise, as thus there is still a logical connection present in the unmet desire (of the pain to go away) and the pain subsisting in the agent.

This being said, the *harm* that we are speaking of in this instance entails another kind of desire being unmet, i.e. an implicit desire[8]. For example, when Engelbert is walking down the street to the baker's, he has an implicit desire not to be punched in the face. However, more specifically, he has the desire not to have the mental state of *oh my god my face is in excruciating physical pain* (Again, if he did not have this deeper desire to avoid this mental state, then being punched in the face should not be a source of displeasure as far as mental states are concerned.). Thus, when Smith socks him one, his

[7] This is of course assuming that the agent in question possesses normative faculties and does not have any anomalous mental characteristics such as being a masochist or a hypochondriac, for instance.

[8] By "implicit desire," I mean a desire such that one does not consciously affirm that there is any such desire until the situation calls for some sort of introspection, upon which there is indeed found to be one. For example, we all have an implicit desire not to get hit in the face (most of us, at least). However, we do not walk amidst ourselves consciously asserting this desire. It is simply a given assumption that this is something we do not want to have occur, and thus it is correct in saying that we do in fact have the desire *not* to get hit in the face. Though normatively speaking, this only comes up in conscious affirmation a) when someone is specifically looking introspectively at ones desires, or b) after this desire has failed to be met. For example: "Smith, why did you hit me in the face?!" "Oh, sorry Jones, I did not think you would mind." "Of course I mind, I certainly did not want to get a shiner!" Hence, we see the implicit desire Jones had was not made elicit until after the fact, but he nevertheless had it.

implicit desire has failed to be satisfied, which consequently results in his mental state of pain. The same can be said for any other account of physical pain. This would then show that the objection of (i) on these grounds is not detrimental to the argument.

With this objection now assuaged, we are thus able to move forward. It is beneficial to remind oneself that unlike the argument put forward by Cicero, this argument has focused on the connection between pain and unmet desires (which is the reason for its success). By virtue of this restatement of pain (which has been defended above), it is simply not possible for one to have any higher pleasures once the absence of pain has been met—seeing as the absence of pain *is exactly* what having all pleasures possible at a given time *is*. Thus, while the previous move from (iii) to (iv) in Cicero's argument was rather perplexing, such a move cogently follows from the restated argument.

V. The Two Lives, Revisited

However, now it must be shown that this new argument explains the *two lives* objection in a satisfactory manner—avoiding the intuitive power that it seemingly wielded against the former argument. For sake of convenience, I shall restate it as follows:

> *In life one, we see that the agent has the absence of pain, but only eats bread and drinks water. In life two, we observe the same absence of pain, but the agent eats filet mignon and drinks wine. Intuitively, the latter seems to experience more pleasure than the former. Certainly, this should show that the absence of pain is not the highest pleasure possible.*

As previously discussed, this observation carries a significant amount of intuitive weight[9]. However, once these two lives are investigated in a more thorough manner, in light of what we deduced in the reformed argument, we will see that this is not the case. From Cicero's testimony (as stated before), we see that the Epicureans had in mind a certain *pleasure calculus*, where one may forego a certain immediate pleasure or endure a certain pain in order to hopefully obtain a greater pleasure in the "long run" (albeit assumed that humans usually do this quite poorly when "calculating" future events). It is by this method that one makes rational decisions concerning one's actions (58-59). In determining one's *aggregate* experience of pleasure at a given time, we may imagine that a given pleasure is assigned some positive value V and its opposite some negative value $-V$.[10] Given our previous argument, we could imagine that this negative value denotes the desire for something (and consequently will be indicative of displeasure, or pain, if the desire is not met), and that positive value denotes the fulfillment of such a pleasure. Thus, when the desire is fulfilled, we are met with an equilibrium mirroring the tranquility experienced when all pain is absent.

With this in mind, let us now turn to the two examples. The opponent wants to assert that because filet mignon and wine are more pleasurable to ingest than bread and water, and *both lives* are absent of pain, then the latter should more pleasurable than the former since there is some apparent extra pleasure value. However, there is a grave

[9] As discussed in lecture by Professor Antonia Lolordo on November 14th, 2007.

misunderstanding that has taken place in this objection. As stated by the reformed argument, one is absent of pain only when all desires have been met. In other words, all negative values in terms of the pleasure calculus have been canceled out (i.e. fulfilled) with the positive ones. So, while the first life may only involve eating bread and drinking water, there is *no* desire for steak and wine. Thus, the first life is in equilibrium. The second life is also in equilibrium even though the agent is eating rather tasty foodstuffs, because before the person was continuously ingesting the steak and wine, there was a desire for such foodstuffs[11]. Thus, the negative value has been canceled out with the positive. Therefore, the two lives are equivalent to each other in terms of the aggregate pleasure value. There is no difference in the aggregate pleasure that each agent experiences in his respective life.

It is a clever move on the opponent's part to mention the steak and wine separately from the fact that all pain is absent, for it gives the illusion of both lives being equal, *plus some given pleasure* for which the steak and wine allow. Yet, as we have shown, this is not the case. In fact, if all pain is absent from the second life, then the steak and wine should not even be mentioned, as it is included in the former fact. Though, perhaps the objector will try to counter the following: *It should be possible for one to experience a pleasure, but not have the desire for it. If this is so, then the two lives would not be equivalent at all.* I certainly agree that *if* this was possible, then the two lives objection would have substantial weight, regardless of its initial misconceptions of aggregate pleasure. However, I do not see this as a viable possibility. To illustrate this point, let us observe a thought experiment:

> *Billy is one of the fortunate fellows who has no pain in his life (because, of course, all of his desires have been met—meaning that he is experiencing the highest pleasure at time t). One day, while walking down the street, a magical fairy appears from thin air and instantaneously erects an honorific statue of Billy on the corner of the street (and we may imagine, for the purposes of this thought experiment, that it is the most beautiful statue Billy has ever seen). Billy, of course, did not have this desire before; for if he did, he would certainly not be some one in which all pain is absent. Thus, it appears (albeit prima facie) that Billy has experienced a pleasure that is free from any desire of it. Does this not show that the theory of pleasure calculus as employed in the two lives is false?*

Indeed, this seems quite perplexing. However, one must ask what would happen if the statue was suddenly taken away. Surely, all would agree that this would cause Billy some sort of pain[12]. However, therein lies the rub. If Billy then suffers pain, it must be because there was a desire somewhere that was not fulfilled. More specifically, there was a desire that arose in the time that the statue appeared and disappeared. But where did this desire come from? It seems that the instant Billy saw the statue, a desire to have the statue *remain* accompanied the initial pleasure that Billy observed. Thus, even though it *appears* that Billy's pleasure (that the statue evoked) was unaccompanied by any desire,

[11] Or, perhaps it resulted when the agent ate for the first time.

[12] This is so, for we have deduced that pleasure comes from desires being fulfilled, and taking away the pleasure will thus leave one with unmet desires, and thus with pain.

the two entities materialize almost instantaneously[13]. Therefore we see, like the other aforementioned instances, that pleasure ensues when the desires are continuously satisfied, hence the equilibrium that is present.

VI. Concerns about the Pleasure Calculus

However, another objection may arise concerning how these arbitrary "pleasure values" are calculated. One might object that this makes an incorrect assumption that *V* and *–V* will have always have the same arbitrary "pleasure value" in terms of absolute value, when clearly it seems that one can conceive of an instance of refutation via thought experiment. Such would be the following: *Gaylord wants lamb chops for dinner (rather, he desires them). Thus, as above, this desire results in the negative pleasure value –V. Now, it just so happens that Vanna White (who happens to love lamb chops) will be walking by Gaylord's house during the time he is cooking the chops. If he cooks them, she will smell them, and will desire to come in and eat them. Being that Gaylord happens to cook the best chops in the neighborhood, Vanna will be so enamored with his cooking skills that she will want to marry him. Thus, on one end (namely, if the desire for lamb chops is unmet) the pleasure value will be –V, but on the other hand (namely that Gaylord will cook the chops and consequently marry Vanna White), Gaylord will reach equilibrium (by obtaining V) AND gain a seemingly insurmountable amount of pleasure. Thus, it seems that in this instance the events denoted by V and –V do not have the same absolute value at all when contemplating the adherence or the neglecting of a given desire.* Such is their objection.

However, it is important to note that this is a causal sequence of events, and the events tied to Gaylord cooking the lamb chops would *not* all constitute V, but rather V (cooking the chops) and some other values of W, X, Y, etc. correlating to having Ms. White come over, etc. Yet, even if we *did* treat all of the causal events as one, it is much like the statue example (see page nine), in that once these events concerning Ms. White have occurred, a desire for them (to continue) develops. Furthermore, as the experience of pleasure increases in magnitude, so too does Gaylord's desire for these pleasures, which continues to counter act each other. Thus, it seems that there really is no imbalance concerning the equilibrium of aggregate pleasure/pain after further inquiry into the matter.

Now that we have reached a better understanding of both the argument and the nature of generating pleasure, we are now able to revisit the Epicurean "advice" of eliminating nonessential desires. However, this time it becomes a matter of simple pragmatism as opposed to something needed to get around the issue that arises from Cicero's rendition of the argument. While there is nothing wrong with having more desires so long as they are met continuously, fulfilling them (and thus being rid of pain)

[13] I say almost, because one may argue that Billy had to experience pleasure before he could discern that there was a desire to keep the statue around (for surely the rational person of normative faculties would not desire something that was not pleasurable). If the statue was rather ugly, then obviously there would be no such desire. So, perhaps Billy felt pleasure for a *very* short time frame where there was only time for Billy to discern that he was experiencing pleasure. Likewise, perhaps at the time that the agent in the "second life" eats steak, there is a moment where the two lives are not equal in their aggregate pleasure value, but it is only until the agent realizes that eating the steak is pleasurable, which most would agree would be negligibly short.

becomes more of a risky endeavor, as one is introducing more and more variables that could go awry. Thus, we may say that while the two different lives are equivalent in terms of the aggregate pleasure experienced, it might be *better* to lead the first, more modest life, since there is less chance of something going awry and consequently resulting in pain. However, this is seemingly the only way that one could be advised to pick one over the other.

VII. Conclusion

Therefore, we find that the refocusing of pain in terms of being unmet desire not only fits in accordance with our own actions, but also frees the Epicurean argument of the troubles it was threatened by earlier in the Ciceronian account. Consequently, the *two lives* objection to the Epicurean thus ceases to have any significant claim, since the second argument for the absence of pain now includes all possible pleasures at *t*. Hence, we find the separation of "absence of pain" and "filet mignon and wine" to be misleading in juxtaposition with the first life, as this distinction between "no pain" and the relative foodstuffs is no longer needed. While this is not presumed to be the "winning point" for the legitimacy of Epicurean ethics, it should serve as a testament that sometimes so much as a simple syntactic re-description illuminates demonstrable information not seen before hand.

Joshua M. Mitchell
University of Virginia

References

Inwood, Brad and Gerson, L.P. The Testimony of Cicero. Hellenistic Philosophy. (Indianapolis/Cambridge: Hackett Company, 1997): 58-59.

LoLordo, Antonia. (Lecture, University of Virginia, Charlottesville, VA, November 14, 2007).