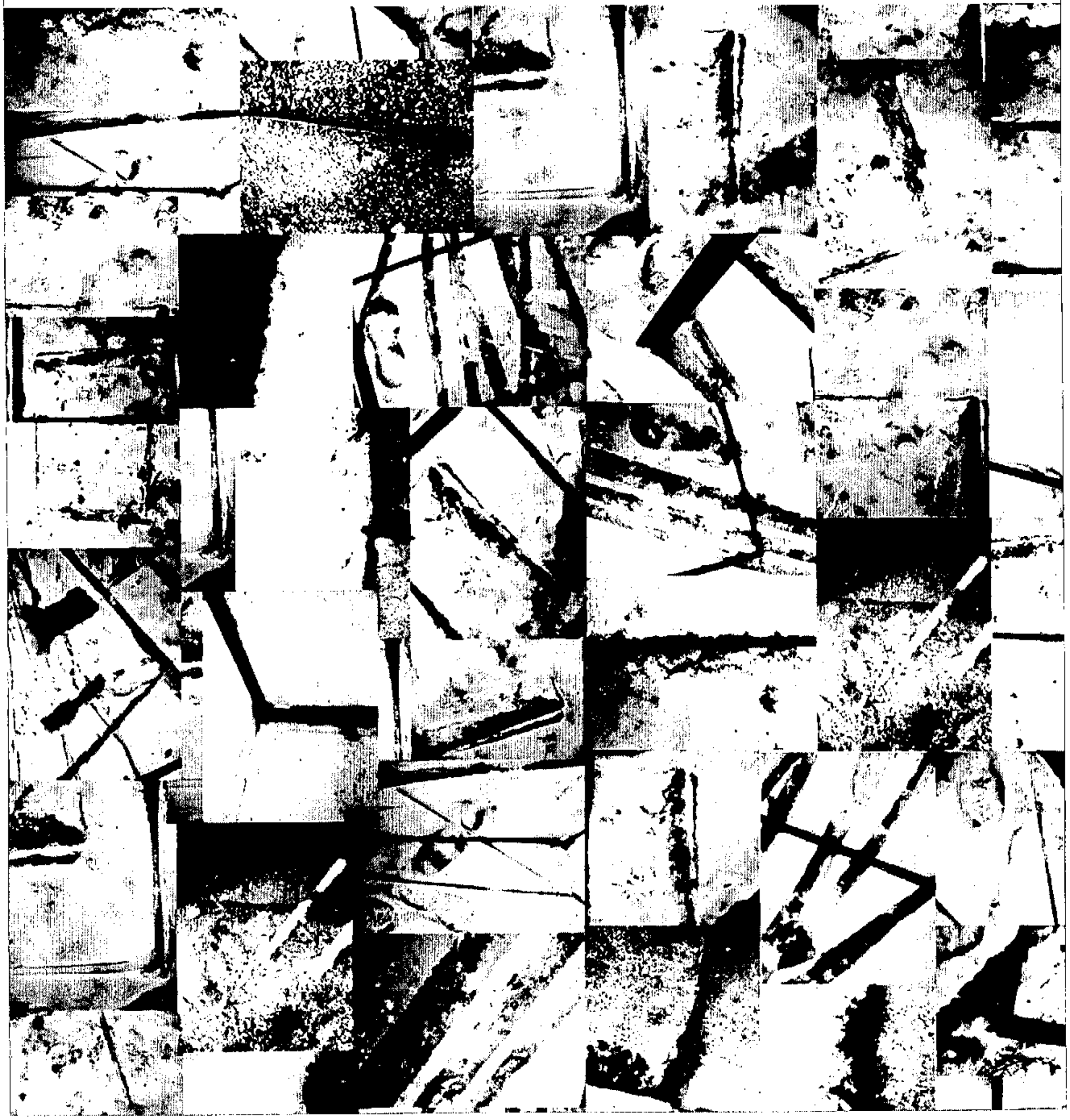
Meteorite

Fall 2007

The Student Journal of Philosophy
at the University of Michigan



Special thanks to the University of Michigan Philosophy department, especially Boris Kment and Ian Proops, Cafe Ambrosia, WCBN, Emigre type foundry, and Alvin Plantinga.

Subscription and submission information can be found at: www.meteoritejournal.com

This journal is funded by the Department of Philosophy at the University of Michigan - Ann Arbor

Copyright inquiries and requests for back issues should be sent to:

Meteorite c/o
Department of Philosophy
2215 Angell Hall
435 S. State St.
Ann Arbor, MI 48109
USA

ISSN: 1099-8764

Meteorite is typeset in Mrs. Eaves, a font designed by the Emigre digital type foundry in Berkeley, California.

Contents

An Interview With Alvin Plantinga

Joshua Blanchard
University of Michigan

Joshua Blanchard: Given that to have warrant a belief must be produced by cognitive faculties in an epistemically friendly environment with a design plan aimed at truth, how do you account for the old problem of knowing that we know? It seems to me that to apply the criterion of warrant, we would have to have some awareness that our faculties are functioning properly.

Alvin Plantinga: That seems right. But I don't know of any general way to tell, for example, when your cognitive faculties are functioning properly. You could be mistaken about that. Still you could be mistaken about most anything, and you still know lots of things. So I'd be inclined to say, right now, that my vision is functioning properly. I can't really give an argument that would satisfy a skeptic, but I don't know if that's necessary in order to know that my vision is functioning properly. And I guess I also think I know that I'm in a friendly cognitive environment. I'm not a brain in a vat, I'm not on some foreign planet where everything goes wrong with cognitive function. And I guess I take it for granted that when I'm functioning properly then for the most part my beliefs would be true. I guess I assume this, just as everyone else does. And I'd also be inclined to think I know it, although that's not something one can sensibly give an argument for. To give an argument for this conclusion, you'd have to be taking for granted that it was true. You'd be taking the conclusion for granted at each step along the way—for example, in proposing a given premise and in seeing the connection between premises and conclusion. So I don't think there's any special problem in knowing that you know, but there isn't any sort of general requirement either – there's no requirement that in order to know, you have to know that you know. Nor is there any general recipe you could give, which is such that if you followed it, you'll know with respect to a given proposition whether you do or don't know it.

JB: So it obviously functions as a definition. Would you then say that you know these things in a more "basic" way?

AP: I'm inclined to say you do. You could have specific kinds of defeaters where the defeater-defeaters were propositions you don't know in the basic way, but ordinarily I think you know in the basic way that your faculties are functioning properly and that for the most part they give you true beliefs. I say this because, if you take it perfectly generally, that's the only way you can possibly know it. You can't know it on the basis of arguments, because these arguments would be epistemically circular. And it seems to me that it's part of our design plan to make these assumptions. If we didn't make them, we'd be in really deep epistemic trouble.

JB: I have a more theological question regarding the Calvin/Aquinas model and the IIHS ("Internal Instigation of the Holy Spirit"), which you modify and utilize in *Warranted Christian Belief.* As far as I understand Karl Barth, he though at least in part that any knowledge of God can only be yielded by a kind of top-down revelation. And it seems like there's some affinity between that and the Holy Spirit requirement of the Calvin/Aquinas model. Or do you think that Natural Theology provides other avenues to gain significant knowledge of God?

AP: We should distinguish two things here. Natural knowledge of God would include natural theology, of course, but there are also other natural ways to know God--ways that don't involve anything supernatural. So Calvin speaks of the *sensus divinitatis* – that would produce natural knowledge of God, although it wouldn't be natural theology. Natural theology would be a matter of providing arguments and so on. If Barth thought that any knowledge of God has to come by something supernatural, that *any* knowledge of God has to come that way, I wouldn't agree with that. I would think, with Calvin, that you can have knowledge of God via the *sensus divinitatis*, and that this is an inbuilt part of human nature. Now when it comes to natural theology and arguments for the existence of God, although I think there are some pretty good arguments –

JB: "Two dozen or so"?

AP: Right, two dozen or so. But I don't think any one of them is or the whole bunch of them together are strong enough to support the sort of belief in God that, say, a serious Christian or Jew or Muslim actually has.

JB: I'm curious, actually – out of the two dozen or so arguments that you summarized in that piece, which is the strongest, if you had to say?

AP: I think the Ontological argument is one of the weaker ones, because the premise and conclusion are so close together. I would say the "Fine Tuning" arguments are quite good arguments. Also, I happen to like the argument from set theory, which goes something like this. There are all these sets. At the bottom level there are nonsets; at the next level, sets of nonsets; at the next level, sets of items at the first two levels, and so on. No set is a member of itself. And sets have their members essentially; if a given member of set S had not existed, S would not have existed either. Now the way of thinking about sets that fits best with these characteristics is Cantor's: a set is really a matter of particulars being *collected*, that is*, thought together* by some mind. That's how I would think of sets. But then if there are all these sets—for example, if there's a set of natural numbers—the mind in question can't be a human mind, or even, I would say a finite mind. No such mind could collect and think together all the natural numbers. So the mind would have to be that of a being of enormously greater intellectual powers than human beings. And the best candidate, I would think, would be God. I think this is a pretty good argument. I also think the moral argument is a good argument. This is the argument that a certain act's being right or wrong essentially involves reference to God in one way or another; and some acts are right or wrong. I think it's a pretty good argument too. But with these and all the other arguments, you can sensibly dispute and reject all of them. For example, you can say that's not the way it is with morality at all. Even in a naturalistic universe, you say, there might very well be right and wrong, good and bad. And with respect to set theoretical arguments, you could just propose some other account of sets, although I don't know what it would be. In the same way, the fine tuning arguments can be opposed. So I think these arguments are pretty good arguments as far as philosophical arguments go, but they don't suffice to support genuine conviction that there is such a person as God.

JB: So do you think that, say, the project of Richard Swinburne is not quite on the mark? He's quite confident in probabilifying first the existnece of God, and second the Resurrection. As you know, he recently said the resurrection was… 94% probable?

AP: I think it was 97%.

JB: 97% probable. That's a pretty strong claim.

AP: Very strong. In his book *The Existence of God*, I think at the end what he claims to have shown is that given our evidence the existence of God is more likely than not.

JB: Right, more than 0.5

AP: More than 0.5. But again, that's pretty slender to actually build a conviction on.

JB: But the Resurrection is 0.97!

AP: That's a little hard to believe. It shouldn't turn out that way, that the probability of the Resurrection is so much higher than the probablility of the existence of God. But I greatly respect his work; it's really excellent work. Still, taking the arguments the way he takes them, again I would say that the way in which I believe in God, or most people I know believe in God, wouldn't be supported or made justified or rational by virtue of an argument where the probability of God is about 0.5. It would have to be much greater than that.

JB: One thing a student brought up at the UM Socratic Club was in reaction to your article "Intellectual Sophistication and Basic Belief in God." You wrote that if the argument from evil has some degree of warrant for a person, but it's of less strength than belief in God, not even an attempt at a defeater-defeater is needed. Somebody suggested – and I don't know if this is a valid conception – that if you added together the warrants of all the atheological arguments, if that would provide a reason for the theist to develop defeater-defeaters or to strengthen her belief in God. So if certain arguments individually have less warrant than the single belief in theism, can the host of them together still constitute an objection?

AP: The only argument that has any real bite or any real promise, it seems to me, is the argument from evil. I don't know what the other ones would be. There's Anthony Kenny who, following Wittgenstein, thinks that only something that has a body could have a mind. But is there any reason to believe that? Why couldn't there be a disembodied mind? I can easily conceive or imagine of myself as disembodied. Richard Dawkins has another argument: The existence of God, he says, is *extremely improbable* because God would have to be incredibly complex in order to have been able to create the world. I don't see the strength to that either. In fact, on the account of complexity Dawkins proposes—having many parts in an arrangement that would be very improbable on chance alone-- God, as theists think of him, would not be complex, because He is not material and hence doesn't literally have parts. Maybe there is some sense in which God is complex—He knows a lot, for example—but why think something complex in *that* sense would have to be improbable? Dawkins mentions another argument that goes like this: God has to be omnipotent, omniscient and wholly good; but if he's omniscient he can't change his mind; and if he can't change his mind, there's something he can't do, so he's not omnipotent after all. Again, not much of an argument. So you can add these anthitheistic arguments together, but I don't think you're going to get much more – maybe nothing more – than the problem of evil.

JB: But do you think warrants could be added together? Is that a sensible conception?

AP: I don't know if you can add warrants together, but you might have several arguments for the same conclusion where, taking them all together is one big argument with several parts – that's a stronger argument than any of the arguments taken individually. In fact, I'd say that's the way it is with the "Two Dozen or so Good theistic arguments."

JB: That's similar to Swinburne's approach.

AP: Right; it's the *cumulative case argument*, as they call it. The same could theoretically go for atheism, except that (as far as I can see) there is only one promising argument.

JB: On basic belief, one of the most common objections, which you do address, is that if belief in God can be basic, then any crazy belief can be basic. There's the a version of the "Great Pumpkin Objection" that I think Michael Martin proposes. And you do address this in *Warranted Christian Belief.* But could you just run through how you respond to this? Specifically, there is the suggestion that the Voodoo epistemologists could come out with a story like the Christian story, etc.

AP: Well the "Great Pumpkin Objection" put as you just put it, is obviously a non-starter. Suppose I think that elementary arithmetic beliefs can be taken as basic. Can't I say the same thing there: "If you can take those as basic, why can't you take *anything* as basic?" But clearly that doesn't make much sense. The mere fact that you say about certain beliefs that they are properly basic obviously doesn't in any way commit you to think that *all* beliefs are properly basic. That objection doesn't go anywhere.

Michael Martin's argument we can call "Son of Great Pumpkin." And it goes like this: Couldn't the Voodoo epistemologists say the same thing? Couldn't she give the same argument about Voodoo that I give about Christian belief? And couldn't the naturalist also say the same thing? And doesn't this show that my argument is mistaken?

But that depends on what you think I was arguing for in that book. My conclusion was that there aren't any *de jure* objections that aren't based in *de facto* objections; there aren't any decent *de jure* objections that don't depend on *de facto* objections to theistic belief. My particular way of arguing for this is such that you might be able to say the same, not just for Christian belief, but for other kinds of theistic belief – Jewish belief, Muslim belief, as well. Maybe so. But you can't say the same for a Voodoo belief, or for a naturalistic belief, because the central part of my argument involved the thought that if Christianity is true, then very likely it is warranted. And a central premise here involves what God would want to do; it involves taking God as an agent who would want us to know about him. None of that's going to work out in the case of naturalism or Voodoo. But it might work out in these other cases I was mentioning. So the consequence would be that the same thing would go for these other theistic belief systems: if they are true they are probably warranted, and you won't be able to find any good *de jure* objections that aren't based in a *de facto* objection. But that's not a problem, as far as I can see. That's not a reason to turn up your nose at my argument. If what I was concluding was that Christian belief is *true* or that it has warrant sufficient for knowledge, then if you could give the same argument for beliefs incompatible with it, that would be a problem. But we don't have that situation at all.

JB: One possible frustration is that with some basic beliefs – like my basic belief that the tree is there – not only do I have a convincing experience, but I also have the experience of other people corroborating that belief and so on. On the other hand, with basic belief in God, even given a very compelling religious experience, there are quite a few dissenters who are additionally epistemic peers in many ways. So imagine only schizophrenics didn't believe in God or something, it wouldn't be a concern. But there is widespread enough disbelief where it seems there's at least something different about belief in God than belief in, say, my mother. So if I believed in my imaginary friend and nobody else believed in it, I would have good reason to doubt my cognitive faculties. Admittedly belief in God is not in quite as dire a situation (since there is a large epistemic community of theists), but it's not as widespread a community that believes in trees.

AP: No, but it's pretty widespread. I would guess nine out of ten people in the world believe in God or something very much like God. It's the atheists and agnostics who are in a small minority and the atheists are in a *tiny* minority. So the theistic community is pretty substantial. And if you include not just the present but go back over history it would be at least that high overall. But there still is that difference, that's right. However, this is not enough to give me a defeater. Consider my philosophical beliefs. For any philosophical belief I hold, there are a lot of people who apparently are my epistemic peers, who disagree with me about that belief. Does that give me a defeater for any philosophical view I've got? If, let's say, 10% of the relevant community disagreeing with me is sufficient for my having a defeater, I wouldn't be able to hold any interesting philosophical beliefs at all!

JB: But it would be *remarkable* if there were as many tree-deniers as there are atheists, wouldn't it? That would be remarkable.

AP: Yes, that would be remarkable. So perhaps 5% of the population of the world are atheists, and not nearly 5% are tree-deniers. There are very few tree-deniers.

JB: So there's some difference. Is belief in God therefore weaker in some sense?

AP: Well the Christian answer, of course, has to do with sin. There are what they used to call the "noetic effects of sin." That was the old Princeton phrase.

JB: That's a chapter in your book.

AP: Right.

JB: "Sin and its Cognitive Consequences."

AP: Right, that's what I'd be inclined to say. We human beings – our minds have been darkened in certain ways as a result of sin. Not only that but our wills have been warped so that many of us don't *want* it to be the case that there is some person as God. I have friends like that. A main obstacle for their being theists is that they didn't want it to be that there is this great being who is privy to your every thought and such that you owe him allegiance and obedience. They think it was a kind of insult to human autonomy that there be such a being. From a Christian perspective that's a result of pride, of sin, and that's one way in which there are cognitive consequences of sin, or noetic effects. And it need not be only in that way. So Christians have always thought that the noetic effects of sin are centered in our knowledge of God and our knowledge and reactions to other people. I would say that this accounts for the difference between the number of tree-deniers and the number of God-deniers.

JB: My last set of questions is on the evolutionary argument against naturalism. First of all, on the scholarly community, how do you find the reception of the argument in general?

AP: It's all over the place. I'm inclined to think that people either like it quite a bit or they really hate it. It's not as if most people are sort of semi-indifferent to it. Naturalists all tend to hate it, and lots of Christians really like it.

JB: Do you think it's convinced anyone against naturalism? Seems doubtful.

AP: I don't think it's convinced any mature naturalist philosophers. But it doesn't have to, in order to be a good argument. In order to be a useful argument, it might convince students. I remember giving this argument at the University of Wisconsin. Later on I heard one student say, "That was the best argument for the existence of God I've ever heard!" And another student I know who was sort of inclined toward being a theist at the time, became a theist and later on a Christian. So to be useful, even with respect to moving people, an argument doesn't have to be such that it moves a full-grown mature naturalist who has been established in naturalism for the last 30 years. But even if it didn't actually move anybody, it could still be a really good argument. First of all it could just be right, and that would be all by itself really good. And it also could encourage theistic and Christian thinkers, it could be a source of encouragement and fit in with other ways of supporting Christian belief and the like.

JB: I was reading one critical review of the argument, by Fitleson and Sober, and they asked some questions about the probability of R [the reliability of our cognitive faculties], and how it gets its warrant. How do you address the objection that R can get its support elsewhere, rather than from E and N [the conjunction of evolution and naturalism]. Say, through some kind of basic belief in R, or through one's experience in the world.

AP: Well, I think we all do believe R in the basic way, and that it has warrant in that way, but that doesn't mean you can't get a defeater for it. I believe R in the basic way, but if I come to think I'm a brain in a vat, I'm going to have a defeater for it. If I come to think I have mad cow disease, or that I'm taking some drug that destroys reliability in four out of five cases, then I get defeaters for it.

JB: But if, as you suggest, it's merely inscrutible on E and N, and R has a great degree of warrant for a person, couldn't they simply believe that something improbable happened on E and N?

AP: Sometimes that works out. But this is a different case. Suppose there's a drug that destroys reliability, suppose I think I've taken it, and suppose I think the probability that somebody's cognitive faculties are reliable given that they've taken the drug is 0.1. The fact that ordinarily R has a great deal of warrant for me doesn't mean that it's perfectly sensible to think, "Well in my case something really unusual happened." That's not a reasonable reaction. It's not as if I have any contrary evidence. As a matter of fact, I couldn't really have evidence for "R" in this case. You have to think about your own case the way you'd think about somebody else's. So suppose I've learned about you, that you've taken this drug and the chances are nine out of ten that you're not reliable. I would not for a moment continue to believe you were reliable. I would certainly have a defeater for that belief, and I don't see how things would be any different in my own case.

JB: Maybe this is not properly analogous, but if I have a belief that a friend of mine is doing well in school and then I find out several facts about his childhood that probabilify that he would not do well in school, I'm not going to abandon my belief that he does well in school.

AP: No, certainly not. So I guess your point would be that there are plenty of things I might believe, on which R is unlikely, that don't constitute a defeater for R. And that will hold in general. It's not the case that if I come to believe X, which is such that the probability of some Y I believe is low on X, that X is automatically a defeater for Y. No, you have to look at these things one at a time. So again take the case where I know that you've taken this drug. I think that is a defeater, even though there might be other things I know with respect to which it's very likely that you *are* reliable. I mean, maybe you're a professor of physics. Maybe you've won the Nobel Prize, as far as that goes. Still, the fact that you've taken this drug trumps these other things. And I think it's the same thing with respect to N and E. These beliefs of yours about how it is that your cognitive faculties got to be the way they are – these are crucially relevant beliefs with respect to whether or not you properly think R is true.

JB: I've noticed that there are two strands, I would say, of what we might call "Christian Philosophy." There's a kind of popular literature written by people like Ravi Zacharias, maybe some by J.P. Moreland and William Lane Craig, etc. These guys publish a lot if literature, whereas there's a more analytic side including Alston, Quinn, Swinburne, yourself, and others, which we might call more "rigorous" work. Do you think this is a positive demarcation? Certainly when you publish a lot of popular literature, it's not necessarily as rigorous, and it also can inspires critics like Dawkins to respond to arguments at their weakest and not at their strongest.

AP: Well, I think it's really important that philosophers not just do the more rigorous kind of thing, extremely important. If you had to choose between the two, I'm not sure which you should choose. It's very important that both be done. And I think too many philosophers – like myself in my early days anyway – value the one kind too much over the other, value the rigorous kind too highly. Partly it's just really fun to work on hard arguments and analyses and the like. Also, its gives you this big feeling of accomplishment when you get something right. You also get this prestige among your peers, whereas if you write popular stuff your peers may look their noses down upon you. But the fact is that for the Christian community, theologians, scientists, and others as well should do what they can to help support and encourage the whole Christian community – not just other philosophers or scientists or historians. So, my hats off to those guys, I'm delighted they do it. People like Bill Craig are perfectly capable of doing both. I've tried to do more of the other kind myself but I'm not all that successful at it. But I do try.

JB: That recent article on Dawkins was pretty good.

AP: Oh, did you like that one?

JB: Yeah. Speaking of popular philosophy, had you read C.S. Lewis' brief critique of Naturalism in his book *Miracles?* People have pointed out that it's very similar to your argument against the theory.

AP: No, I hadn't actually read his argument, and yes, people have pointed that out to me too. But it's not quite all that similar. He's talking about determinism there. And he says if determinism is true, then I can't be confident in any of my beliefs. Or, putting it my way, my believing determinism is true would be a defeater for the idea that my cognitive faculties are reliable. But I don't think that's right. Suppose I thought they were determined, but determined by God. In fact, I think they *are* to a large degree determined. Or at least, if not determined, there certainly is a lot of strong inclination to accept them. It would be *really hard* for me to not believe that there's a book here or that I'm talking to a person. I don't know if it's quite determinism but it's in the neighborhood. With respect to determinism, then, what matters is what you think the ultimate causes of your belief are. If the ultimate cause is a God who has designed us in a certain way to resemble Him, other things in having knowledge, that's not a defeater at all. So I don't think Lewis is quite right on that point. He's right that Naturalism offers a defeater, but it's not via determinism.

JB: You gave an address at the turn of the millennium on the state of Christian philosophy. There has certainly been an increase in the latter half of the 20th century. Do you think the situation looks good? One thing I've noticed is that, at least in the English-speaking world, there is certainly division between secular and religious philosophy, although I don't know what the situation is in the European academy.

AP: Right; there is a clear and obvious division between secular and non-secular, secular and Christian, or secular and more broadly theistic approaches. This division is more obvious and more vigorous that it was, say, 50 years ago. That's in part, I think, because there are far more theistically-inclined philosophers now in the US then there were then. Fifty years ago there were surely some Christian philosophers around, though not nearly as many as now. And the ones that were around were for the most part rather close-mouthed about it. It wasn't part of the public philosophical community; and many philosophers thought that being a Christian and being a philosopher were mutually exclusive. It wasn't a major stream in the community the way it is now. That's one big difference and a very important one.

JB: So you think it's gained respectability?

AP: Respectability? Respectability is very much in the eye of the beholder. There is a recent spate of books by atheist philosophers, who don't think Christian philosophy is respectable; but many others do think it is. In any event, it certainly is part of the mainstream now. There are journals devoted to it like *Philosophia Christi* and *Faith and Philosophy*; and several others. Articles which presuppose Christian belief or at any rate take it seriously are published in other journals as well. In fact Quentin Smith, himself no friend of Christian philosophy, laments, in a recent issue of *Philo* that philosophy has become desecularized. This is clearly a significant change from fifty years ago.

Believe It: The King of France Still Reigns

Adam Rigoni
University of Michigan

Introduction

In this paper I will examine and criticize Von Fintel's theory in "*Would you Believe It? The King of France is Back*,[1] which attempts to explain many people's truth-value intuitions vis a vis sentences containing non-referring definite descriptions. First, I will provide background information needed to understand what motivated Von Fintel's account. Second, I will reconstruct Von Fintel's theory. Third, I will put forward my objection to Von Fintel's explanation of the independence of counter evidence and offer an addition to Von Fintel's theory that will solve this problem.

1: Background

The truth-value determinacy of sentences containing non-referring definite descriptions[2], such as

(1) The King of France is bald.
(2) The King of France is wise.[3]

has long been disputed amongst philosophers of language. Russell famously thought that the truth conditions for such a sentence, or more precisely, the truth conditions for the propositions expressed by such a sentence are (i) there exists one, and only one king of France and (ii) that one is bald[4]. Or more formally, if "b(x)" is a predicate meaning "x is bald" and "k(x)" is a predicate meaning "x is the king of France" then "The king of France is bald," expresses this proposition: >x[k(x) & œ(y)[k(y) e (y=x)] & b(x)] .

Russell therefore held that "The king of France is bald," expresses a false proposition[5] as it fails to mean the first truth condition, because there does not exist a king of France. For Russell, truth-value gaps are to be avoided as they violate the law of excluded middle: By the law of excluded middle, either 'A is B' or 'A is not B' must be true. Hence either' the present King of France is bald' or 'the present King of France is not bald' must betrue.[6]

Russell is here criticizing Frege's view that such sentences are truth-value indeterminate nonsense. He writes, "One would suppose [on Frege's view] that 'the King of France is bald,' ought to be nonsense; but it is not nonsense since it is plainly false [emphasis added]."[7] Here Russell's appeal to truth-value intuition, emphasized in the quotations, is manifest.

Strawson disagrees with Russell's intuition and analysis. He argues, in On Referring, that it is not sentences, but uses of sentences by a speaker in a conversational context, that are true or false. To use (1) or (2) in a

[1] Henceforth abbreviated as *Would You Believe It?*

[2] Here, and throughout the paper, I ignore phrases like "The whale" in the "The whale is a mammal" where we have a generic use of the definite description. These are largely considered a separate issue and have no bearing on my project in this paper.

[3] Somewhere in the dialectic between Russell and Strawson the example used changed from (1) to (2). Both are essentially the same with regards to the characteristics relevant in this paper, but I've listed both for ease of reference.

[4] See *Principia Mathematica,* 67-68; *Introduction to Mathematical Philosophy,* 167-180

[5] Later in the paper I write of sentences *being* true or false as opposed to *expressing* true or false propositions. Nothing I intend to discuss hinges on this, and a strict Russellian may substitute the latter for the former if he is so inclined.

[6] Russell, *On Denoting*.

[7] *Ibid.*

conversational context, certain background facts must obtain, namely, that the King of France exists. These are the facts that are implied or presupposed by the use of the sentence. But the absence of these background facts does not make the sentence false, as Russell believes. What happens instead is that the question of falsity never arises and the sentence is truth-value indeterminate.

Strawson writes:

> [Using (2) is] to imply that there is a King of France. But this is a very special and odd sense of 'imply'. 'Implies' in this sense is certainly not equivalent to 'entails'(or 'logically implies'). And this comes out from the fact that when, in response to this statement [i.e. (1)], we say (as we should) 'There is no king of France', we should certainly not say we were contradicting the statement that the King of France is wise. We are certainly not saying that it is false. We are, rather, giving a reason for saying that the question of whether it is true or false simply does not arise.[8] (emphasis Strawson's)

Here Strawson is plainly denying Russell's assertion that (1) is plainly false. For clarity of terminology it should be noted that Strawson is here using "implies," but later,

Strawson (1952, 1954) does use the term presupposition and defines it as Frege (1892/1970:69) does: A presupposes B iff A is neither true nor false unless B is true. This has come to be known as the semantic conception of presupposition.[9] Yet Strawson does not deny Russell's assertion completely. He allows that some uses of non-referring definite descriptions do generate intuitively false sentences. He gives the examples along the same lines as these:

(3) My friend went for a drive with the King of France.

(4) The Exhibition was visited yesterday by the King of France.[10]

So now there are sentences using non-referring definite descriptions that are clearly judged false, e.g. (3) and (4), and others about which Russell and Strawson have conflicting intuitions regarding their truth-value, e.g. (1) and (2). Let us call the former "clearly false" and the latter "squeamish" following Von Fintel and Strawson, because we are "squeamish" about assigning a truth-value.

This gives rise to the question of whether truth-value intuitions are of any use in determining the semantic status of these phrases. Strawson clearly thinks that in the case of (2) our intuition is one of a truth-value gap. He writes:

> Suppose he [who uttered (2)] went on to ask you whether you thought that what he had just said was true, or was false...I think you would be inclined, with some hesitation, to say that you didn't do either.[11]

Yet Russell denies this same intuition, writing:

> Suppose, for example, that in some country there was a law that no person could hold public office if he considered it false that the Ruler of the Universe is wise. I think an avowed atheist who took advantage of Mr. Strawson's doctrine to say that he did not hold this proposition false would be regarded as a somewhat shifty character. (Russell 1959: 243-4)[12]

Strawson and Russell are not alone in their intuitions. Von Fintel clearly sides with Strawson in Would You Believe It?, while I have largely Russelian intuitions. It seems that

[8] Strawsoan, *On Referring*, 345-346

[9] Reimer and Bezuidenhout, *Descriptions and Beyond*, 262

[10] *Ibid*., 263

[11] Strawson, *On Referring*, 345

[12] Quoted in Von Fintel, *Would You Believe It?*, 273

direct truth-value intuitions just cannot help us sort this matter out. This is the conclusion that Von Fintel (Would You Believe It?), Soames (1976: 169) and Thomason (1990:327)[13] have reached as well.

The next candidate for deciding this matter was based on examining presuppositions and how they behave, with a focus on existence presuppositions. It should be noted, as Bach has pointed out, that if "presupposition" is used in the semantic sense, as it is used by Strawson, i.e. A presupposes B iff A is neither true nor false unless B is true, then we have already assumed Russell is incorrect.[14] Even to use presupposition in a loose sense, as the information needed to felicitously use a phrase in conversation, begs the question against Russell because, for him, if there is no object referred to by a definite description, the sentence could be used as felicitously as any other false sentence. What we might want to say is that by "existence presupposition" we mean the condition that the referent of a definite description exists; this is a truth condition for Russell, but a semantic presupposition for Strawson and Von Fintel. Another way of putting this is that for Russell, (1) asserts both that (i) there is a king of France and (ii) he is bald, while for Strawson (1) presupposes (i) there is a king of France and asserts that (ii) he is bald. The point is that for Russell (i) and (ii) should exhibit the same behavior.

Russell's intuition loses plausibility when scrutinized under Von Fintel's "Hey, wait a minute test." The "Hey, wait a minute test" is one Von Fintel uses to determine what the conversational presuppositions are. For example:

(4) The King of France drives a Mercedes.If a speaker, A, uttered (4) to a listener, B, B might legitimately object:

(4#) Hey, wait a minute. I had no idea France was still a monarchy.
But B could not legitimately object:

(4') Hey, wait a minute. I had no idea he drove a German car!

Hence the dialogue could go:

A: The King of France drives a Mercedes.
B: Hey, wait a minute. I had no idea France was still a monarchy

But it could not go:

A: The King of France drives a Mercedes.
B: Hey, wait a minute. I had no idea he drove a German car.

Now we can see that (i) behaves differently than (ii) in conversation.

This criticism gets even stronger when one considers the projection of presuppositions. The projection of presuppositions occurs when the presuppositions of a phrase containing a definite description stay with it when it is embedded within a larger sentence structure, while the assertions made by the phrase are not carried over. Consider:

(5). I hope that the king of France is bald. One could, in conversation, reasonably criticize me for not making sense by saying:

(#5) Hey, wait a minute. I had no idea that France was still a monarchy.
But it would not be reasonable for someone to criticize me by saying:

(5'). Hey, wait a minute. I had no idea that he was wise.

[13] For the relevant Soames and Thomason passages, see Von Fintel, *Would You Believe It?*, 274

[14] Reimer and Bezuidenhout, *Descriptions and Beyond*, 263

When you hope that X, where X is an embedded clause, you are obviously not asserting X, so it is illegitimate to complain that the truth of X is unknown to you (or that you believed it to be false[15]). Hence (5') would be a bizarre objection to make. But (5#) is a legitimate objection because (i) is relevant in making sense of the sentence. Hence (i) differs from (ii) in that (i) is relevant to understanding (5), i.e. it is projected, while (ii) is not. Moreover, (i) is relevant in cases in which the truth of X is not, which suggests that (i) has a relevance different from a truth condition.

This objection can be evaded while maintaining that there are no truth-value gaps. The objection is pragmatically, i.e. conversationally, based, so Russell and like-minded philosophers could merely say that what makes sense in a conversation is not a good guide to actual truth-values[16]. As Von Fintel notes, there are analyses, e.g. Karttunen and Peters' system (see: Kartutunen and Peters 1979), that assume a Russellian two-value semantics and in which pragmatic presuppositions are encoded at an independent level (Would you Believe It?, p.271).

Thus, it is very odd that Von Fintel grounds his assumption of truth-value indeterminate, i.e. 'gappy', semantics on the basis of how well it functions as a basis for a theory of presupposition behavior. He writes:

> One says that a sentence has the semantic presupposition that p iff the proposition it expresses does not assign a truth-value to the states of affairs where p does not hold. I will work with a Frege-Strawson semantics for definites:
>
> (6)[17] The P is Q expresses a partial proposition which is defined only for worlds in which there is a unique P and which is true only in a world w if the unique P in w is Q in w.
>
> ...This semantics is not one that we can argue for on the basis of raw truth-value intuitions. Rather it's advantages lie in how well it can be used to derive the pragmatic facts about presupposition.[18]

But in the note cited above, he acknowledges that we could have Russellian semantics with an independent level at which pragmatic facts about presupposition can be encoded. There is no reason Von Fintel's presupposition theory could not be encoded at that level and therefore no reason his theory could not be used within a (two-leveled) system of Russellian semantics. Perhaps his claim is that we need a system to deal with presuppositions, which differ from truth-conditions, at some level and that a pure, single-level Russellian system will not allow us to.[19] Yet this is a weaker claim, and it seems that Von Fintel's theory does not accomplish as much as he would like.

Thus the semantic status of phrases like (1) and (2) is still very much up in the air, but another, perhaps more easily answered, question remains: Can we systematically explain the varying truth value intuitions philosophers such as Strawson had about sentences like (3) and (4)? Is there some underlying mechanism at work and, if so, what is it? It is this question that Von Fintel's theory purports to answer directly and, I think, succeeds.

[15] For example, B is not a legitimate objection in this case:
A. I hope that my girlfriend is happy.
B. Hey, wait a minute. Your girlfriend is not happy.

[16] I do not here wish to engage in a debate about how committed Russell was to base his semantics off of practical concerns. It is enough that his two-value semantics could be maintained in spite of the objection.

[17] The numbering here is Von Fintel's, not mine. It has no bearing on the numbering in this paper.

[18] Von Fintel, *Would You Believe It?*, 272

[19] I will not pursue this objection any further because it is ancillary to the thrust of Von Fintel's project, as I see it, which is discussed in the proceeding paragraph.

III: Von Fintel's Project

Von Fintel's main goal is to create a theory that explains why most people think a sentence like:(

> 3) My friend went for a drive with the king of France. is false, while they are squeamish about saying a sentence like (1) The King of France is bald. is false. Note that we are here talking about truth-value judgments, which may or may not be helpful guides to actual truth-value. Von Fintel's concern here is with why people reject some sentences with non-referring definite descriptions outright as false but are squeamish about others.

Before Von Fintel constructs his theory, however, he shows, as he should, why previous analyses are incorrect. One tempting way of explaining people's truth-value judgments is to say that they correspond to the presence or absence of existence presuppositions[20].

The idea is that sentences about which we are squeamish, like (1), have existence presuppositions, while sentences we reject as clearly false, like (3), do not. The theory goes on to claim that we can see which sentences carry existence presuppositions by looking at whether the definite description is in the topic position or in the focus position. If the definite description is in the topic position, then there is an existence presupposition, the failure of which results in us feeling squeamish about assigning it a truth-value. If the definite description is in the focus position, then there is no existence presupposition, and we then judge the sentence clearly false when the presupposition fails to obtain.

It will be helpful[21] to clarify just what the topic-focus distinction is. The topic of a sentence refers to information that can be considered background information, as it is already common ground between the participants in a conversation.[22] The focus of a sentence refers to the new information presented in the sentence. A clear explanation is given by Reimer and Bezuidhout:

> "In English, for example, the subject expression is usually, though by no means invariably used to mark the topic. The predicate expression is then used to make some sort of comment on the topic [the focus aspect of the distinction is sometimes called the "comment" instead of a "focus"]. It-clefting is another device for indicating [the focus aspect of] topic-focus structure. 'It was Mary who visited London' and 'It was London that Mary visited'... are associated with different information structures. In the former case it is already established that someone went to London, and the new information being asserted is that Mary was that person. In the latter case, it is already established that Mary visited some place, and the new information being asserted is that London is that place.[23]

To their examples I add one that does not involve it-clefting:

(6) Adam is hungry.

Here "Adam" is in the topic position, as he is part of the background while the new information, i.e. what is in the focus position, is that he is hungry. Substituting a non-referring definite description in for "Adam" we get a sentence in which the definite description is in topic position, hence it carries and existence presupposition that fails, and thus it makes us squeamish to assign a truth-value to it:

[20] Von Fintel notes (*Would You Believe It?*, 277) that this theory is propounded by Reinhart (1981, 1995); Hajicova (1984); Gundel (1977); Horn (1986); Lambrecht (1994); Erteschik-Shir (1997); and Zubizaretta (198)

[21] It is also necessary for those who, like myself before reading *Would you Believe it?*, have never heard of the distinction before,

[22] Reimer and Bezuidenhout, *Descriptions and Beyond*, 264

[23] *Ibid.*, 264

(7) The King of France is hungry.

This theory accurately predicts our truth-value intuitions in some cases. In (1) "the King of France" is in the topic position and thus there is an existence presupposition we have squeamish feelings about. Yet (3) seems less clear. It appears that the definite description is in the focus position, and the common truth-intuition is that it is clearly false. But does it really lack an existence presupposition?

This is where Von Fintel's objection comes in. Even granting that (3) lacks an existence presupposition which accounts for our rejecting it as clearly false, he has other examples that most would reject as clearly false, but which have an existence presupposition[24]:

(8) A: The king of France attended the APEC conference this week.[25]

B: Hey, wait a minute– I had no idea France is still a monarchy.
B': # Hey, wait a minute– I had no idea that he was at that conference.[26]

Here A makes a claim that most would reject as clearly false. But B legitimately criticizes it by pointing out the lack of a commonly acknowledged referent for the definite description, "the King of France." B', on the other hand, illegitimately criticizes him for asserting information that is not common knowledge between the two of them. Because B's objection is legitimate, (8) has an existence presupposition. But this contradicts the topic/focus theory.

Von Fintel has counterexamples for cases in which the sentence containing the non-referring definite description is embedded in a larger construction, including:

(9) I hope that the king of France attended the APEC conference this week.[27]
One could, in conversation, legitimately object to (9) by saying
(9') Hey, wait a minute– I had no idea France was a monarchy.

Hence, by the "Hey, wait a minute" test, (9) has an existence presupposition as well, despite the fact that the non-referring definite description occurs in an embedded clause.

Yet recall the previous discussion of (3). Recall that it was uncertain whether or not (3) had an existence presupposition, but was clear that the non-referring definite description was in the focus position. Perhaps the topic-focus distinction by itself, without any corresponding presupposition information, is enough to explain our truth-value intuitions.
Alas, Von Fintel has a counter example to this as well:

(10) I had breakfast with the king of France this morning. He and I both had scrambled eggs.[28]

In the second sentence of (10) the definite description, or the pronoun that goes in for

[24] It should be noted that this objection only works if one accepts that the "Hey, wait a minute" test accurately indicates which sentences have presuppositions. One could deny the test, but then he would be in a position to have to explain why the absence of certain conditions, which could not be presuppositions nor truth conditions, can be criticized in this way in conversation. An extended discussion of the merits of Von Fintel's test would be long and tangential at this point, I will assume its accuracy.

[25] I have altered Von Fintel's numbering here and in proceeding numbered sentences so it corresponds to the numbering of sentences in this paper. Cf. footnote 9.

[26] Von Fintel, *Would You Believe It?*, 277

[27] *Ibid*., 277

[28] *Ibid*., 277 (Von Fintel's superscript "F"s have been omitted because I am not using that notation)

it, is in the topic position, but the sentence is still rejected as clearly false.[29] Thus, the topic-focus distinction is utterly useless in helping us explain people's truth-value intuitions. Von Fintel then turns to an analysis given by Lasersohn, upon which Von Fintel's own analysis is closely based. Von Fintel presents what is essentially Lasersohn's theory, but within a simpler framework.[30] I will not bother going through all the reconstructions and revisions Von Fintel does in examining the theory. Instead I will outline it and present its final form using the epistemic revision that completes the reconstruction.

The idea is that the we reject sentences with non-referring definite descriptions as clearly false when we can assign the truth-value independently of knowledge about the non-existence of the relevant referent. Lasersohn writes:

> [a statement of the aforementioned sort] can...be judged false, provided the context makes it possible to determine that the statement could not possibly be true regardless of whether the term has reference or not... Why is it that someone who points at an empty chair and says The King of France is sitting in the chair seems to be saying something false? I would like to suggest it is because even if we suspend our knowledge that there is no King of France, there is no way of consistently extending our information to include the proposition that the King of France is sitting in that chair. Such an extension is impossible because we know the chairto be empty. In contrast, if we suspend our know-ledge that there is no King of France, our information may then be extended to include the proposition that the King of France is bald (1993:115).[31]

A simpler way of understanding this is to think of it in terms of even if -conditionals. Consider:

(11) Even if there is a king of France
(which there isn't), he is still not bald.

(12) Even if there is a king of France
(which there isn't), he is still not sitting in that chair/that chair is still empty.[32]

We are likely to assent to (12) but not to (11) and hence we judge "the king of France is sitting in that chair" to be clearly false and feel squeamish about "the king of France is bald."

Von Fintel formalizes this system into a rule for rejection of a sentence as clearly false[33]. I will here have to introduce some of his notion because Von Fintel distinguishes between what he calls pragmatic truth and falsity, which he denotes by "TRUTH" and "FALSITY," and semantic truth and falsity, which he denotes by "1" and "0".[34] For Von Fintel this is important because he is assuming Frege/Strawson semantics and hence all sentences with non-referring definite descriptions are semantically truth value indeterminate, even those we would reject as clearly/pragmatically false. For Von Fintel some FALSE sentences are semantically neither true nor false. For a Russellian, FALSE sentences are merely a subset of semantically false sentences, i.e. those assigned "0." I find this use of numbers and capitalized letters confusing so I will simply write "pragmatically

[29] Or so Von Fintel claims. I'm not sure that the second sentence is any more clearly false than (1), but there is no use nitpicking about intuitions and Von Fintel has apparently done some empirical research (Von Fintel, *Would You Believe It?*, 293) to support some of his claims, although not this one in particular.

[30] Von Fintel, *Would You Believe It?*, 280

[31] *Ibid.*, 280

[32] *Ibid.*, 283

[33] Bearing in mind that is a system for pragmatic, conversational rejection. Just because we do not reject a sentence as clearly false, does not touch on whether it is semantically false. We could still have the semantics independent of the pragmatics *a´ la* Karttunen and Peters' system (Kartutunen and Peters 1979). See pages 7-8 of this paper.

[34] Von Fintel, *Would You Believe It?*, 280

false" where Von Fintel would write "FALSE" and "pragmatically true" where Von Fintel would write "TRUE." Accordingly, I will write "semantically true" where he would write "1" and "semantically false" where he would write "0."

More notation is needed to get to the formalization. Let D stand for a given body of information, modeled as a consistent set of propositions. We accept sentences as true or reject them as false with respect to this body of information. Let S stand for a sentence. Let p stand for a proposition that is presupposed by S, e.g. "the King of France exists." We also need an epistemic revision process that goes as follows:

> Common-sense epistemic revision
> Remove ¬p from D. Remove any proposition from D that is incompatible with p. Remove any proposition from D that was in D just because ¬p was in D. Add p to D. Close under logical consequence.[35] Let D* be the result of revising D in the above way. The procedure for rejection as pragmatically false is then:

Rejection

> Reject S as pragmatically false with respect to D if and only if for all worlds w compatible with D*, S is semantically false.
>
> Note that in any world compatible with D*, S will have a definite truth value because its presuppositions are fulfilled in D*, by definition of D*.

Lasersohn's theory will accurately predict some of our intuitions. For example:

(13) The king of France is sitting in the chair next to me.

is rejected as pragmatically false because the information (in proposition form) that the chair next to me is empty remains in D*. Conversely, (1) is not rejected as pragmatically false, because the information that the king of France is not bald is not available in D*. With (1) we don't reject it, but we certainly can't accept it, and hence we feel squeamish about it.

Unfortunately, there are counter-examples to this theory as well. Von Fintel offers two examples:

(14) The King of France is on a state visit to Australia this week.

(15) (Coming across an abandoned cell phone on a park bench) This cell phone was left here by the king of France.[36]

Most people reject (14) and (15) as pragmatically false, but under Lasersohn's theory they should be squeamish about them, because our reason for believing the king of France is not in Australia and that he did not leave the cell phone is that the king of France does not exist.[37] (15) is especially damaging because it fails Lasersohn's even-if conditional test too:

(16) Even if there is a king of France (which there isn't), this cell phone was left by someone else.[38]

We would not accept (16) as being pragmatically true (nor semantically true for that matter).

[35] *Ibid.*, 283

[36] *Ibid.*, 285

[37] More specifically in the cell phone case, that no king of France existed when cell phones did.

[38] Von Fintel, *Would You Believe It?*, 285

The question then is, what differentiates (1) from (14)? Von Fintel's insight, upon which his theory is based, is this:

> In the case of (16) but not of (1), there is a contextually salient entity whose properties (known or not known) are in principle enough to falsify the sentence. In (1) there is no contextually salient entity mentioned (other than the king of France) whose properties could establish that (1) is false. In (16), Australia is made salient and can thus furnish an independent foothold for falsification.[39]

In other words, we can make use of information, the only basis for which is that there is no king of France, if we could, in principle, obtain that same information independent of the king's non-existence, from a source that is contextually salient. This is formalized in a modification of the procedure for epistemic revision found on page 14:

> Von Fintel's Conversational Revision
> Remove ¬p from D.
> Remove any proposition from D that is incompatible with p.
> Remove any proposition from D that was in D just because ¬p was in D, unless it could be shown to be true by examining the intrinsic properties of a contextually salient entity.
> Add p to D. Close under logical consequence.[40]

Call the result of revising D in the above manner D!. The procedure for rejecting a sentence as pragmatically false is then:
Rejection

> Reject S as pragmatically false with respect to D if and only if for all worlds w compatible with D!, S is semantically false.

Von Fintel does not specify "what counts as a contextually salient entity of the right kind and what exactly it means to say that the intrinsic properties of that entity are enough to falsify the sentence."[41] Von Fintel says, as an approximation, that "contextually salient entities will be those mentioned in the sentence" and that he will not have much else to say about this (Ibid). Herein lies a problem, but I want to ignore this for now. It is important, first, to see how the theory works, for I think it is essentially correct and it works well.

The theory can explain, for example, why (1) makes us squeamish but sentences like the following do not:

(17) Among the bald people in the world is the king of France.
(18) The king of France is one of the bald people in the world.

(1) does not make salient, nor even mention, the set of bald people in the world. Von Fintel notes, "the predicate bald is not even a referring expression" (Von Fintel, Would you Believe it?, p.286). Conversely (17) and (18) do make such an entity salient, and on the basis of the king of France not being a member of the set of all bald people in the world, we can reject them as pragmatically false.

For Von Fintel, a contextually salient entity does not have to be an entity of the usual kind. He gives examples of claims about particular episodes, which most would reject as pragmatically false, such as:

[39] *Ibid.*, 286
[40] *Ibid.*
[41] *Ibid.*

(19) The king of France is jogging now.[42]
I suppose the idea is that this makes salient the set of events happening right now. The properties that count as intrinsic are also broad, as Von Fintel uses this as an example of a sentence we would reject as pragmatically false that his theory accounts for:

(20) The king of France owns this pen.[43]
Apparently previous owners falls under the intrinsic properties of the pen in question. The rejection of (15) as pragmatically false would be explained in the same way.[44]

Von Fintel further goes on to claim that the contextually salient entity need not be mentioned in any sentence. He writes:

> The contextually salient entity may not have to be mentioned in the same (or any) sentence. David Pesetsky (pers. comm.) reports that The King of France is bald can be judged false [i.e. rejected as pragmatically false] if made in the presence of a list that enumerate all the reigning monarchs of the world together with their hairstyle. (Von Fintel, Would you Believe it?, p.287)
>
> This claim is important to bear in mind because he unknowingly contradicts it later and my objection is based on a similar thought.

So far, so relatively good. But Von Fintel has to explain how independent the counter evidence has to be from the failed presupposition. Von Fintel[45] points out that (1) and (21) The man who Sandy went out with last night is bald [assuming there is noman Sandy went out with last night][46].

Both make salient entities whose intrinsic properties could falsify the sentence, i.e. France and Sandy, respectively, yet both, he claims, make us squeamish. Compare these sentences with (14), which we reject as pragmatically false:

(1) The King of France is bald.
Counter Evidence: France does not have a bald king

(21) The man who Sandy went out with last night is bald [assuming there is no man Sandy went out with last night].
Counter Evidence: Sandy did not go out with any man last night.

(14) The king of France is on a state visit to Australia this week.

Counter Evidence: Australia is not being visited by the king of France this week.

In each case, the only reason we have to believe the counter evidence to be true is that the existence presupposition of the definite description fails. Von Fintel's solution is this:

The counter-evidence in (14) is in principle epistemically independent of the offending [read: non-obtaining] presupposition. While we believe it to be true just because we believe the presupposition to be false, we could conceivably show it to be true while not showing the presupposition to be false.

[42] *Ibid.*, 287

[43] *Ibid.*

[44] There is a problem for Von Fintel in that he later uses the example of "The King of France heard about the car accident on the turnpike last night" (Von Fintel, *Would you Believe it?*, 289) and claims that we feel squeamish about it because the intrinsic properties of the car crash do not include who heard about it. Yet if the intrinsic properties of "this pen" extend to past ownership, then it seems arbitrary that the intrinsic properties of "the car accident" do not extend to who heard about it. I wish to ignore this here, as I frankly do not have an explanation nor solution for it.

[45] Von Fintel, *Would You Believe It?*, 287

[46] I am not at all convinced that this example makes us squeamish. In fact, I think it can be rejected as false. But I will grant Von Fintel that it does for now.

We could travel to Australia and see what's going on.

In contrast, the potential counter evidence in (1) [and (21)] is not epistemically independent of the non-existence of the king of France [or the man who went out with Sandy], even in principle. As soon as we show that France does not have a bald king, we will have show that France does not have a king at all.[47]

But this solution fails and the claims it is based on are simply wrong.

IV: The Objection and Solutions

Before I demonstrate why Von Fintel is wrong, it should be pointed out that if his analysis is correct, then he has still flatly contradicted a claim he made just three pages earlier in the article. If the counter-evidence must be able to, in principle, falsify the sentence without showing the presupposition to be false, then David Pesetsky's list that enumerates all the reigning monarchs of the world together with their hairstyle[48] would not be acceptable counter-evidence. For if we looked at such a list, we would see that the king of France was not on it and thus would have shown that he does not exist, i.e. that the presupposition has failed.

Von Fintel is wrong here because his claim that we could not show, by examining France, that the king of France is not bald without showing that France does not have a king, is false. We must distinguish between "showing" and "examining." Von Fintel sometimes writes of getting counter evidence by showing it to be true, as if "showing" were some activity in the world, e.g. "We travel to Australia and see what's going on."[49] But at other times, when he is writing his formal rules, he writes of getting counter evidence by "examining" intrinsic properties, where "examining" seems more like an abstract mental exercise.[50] My objection is that whether we "show" the evidence to be true or get the evidence from "examining" properties his claim is still false.

We could "show" that the King of France is not bald by going to France and seeing what is going on, without showing that there is no King of France. We could see a poster with pictures of every king of France, without dates of their reigns, and notice that there is no bald guy in the picture. The same information could be represented on a list, much like Petesky's, where every king of France (without dates) and their hairstyle is listed. The evidence could be more manageable: It could be a picture entitled "The Last 5 Kings of France" or one of "All Kings of France since 1700", or the same information on a list. It could even be a picture of the most recent King of France[51]. If we don't see a bald man on any of these pictures or lists, we would have evidence to falsify the assertion that the King of France is bald. We would not, however, have shown that there is no King of France. So, on Von Fintel's theory, we should not be squeamish in this case, but we clearly are not.

[47] Von Fintel, *Would You Believe It?*, 290

[48] *Ibid.*, 287

[49] *Ibid.*, 290

[50] *Ibid*, 286, 290

[51] It might be objected that in some of these cases, namely ones with the most recent king, we would be falsifying based on a false assumption. That is, we would be falsifying on the assumption that the most recent king is the current king. I am not sure that this weakens the objection because I'm not certain if that assertion is false given that the king of France exists. In any case, we could read a plaque, perhaps at a museum, that states that France has never had a bald king. Now this leaves no room for incorrect identity statements, but still shows Von Fintel to be mistaken.

But perhaps the whole idea of "showing" as some action we undertake is bizarre and the "real" analysis deals with "examining." Yet an examination of the intrinsic properties of France could still show that there is no bald king, without showing that there is no king at all. I will suppose that the properties we examine are in the form of propositions. Any of the following propositions will suffice:

P1. France has never had a bald king.
P2. None of the last 5 kings of France has been bald.
P3. Since 1700, no king of France has been bald.

Now it may be objected that these propositions are not "intrinsic." But on what grounds is this based? If propositions identifying past owners are intrinsic properties of a pen, then I see no reason why these are not intrinsic properties of France. The Sandy example, (21), can be dealt with the same way:

P4. Sandy never dates bald men.

One way of getting out of this is to claim that we cannot just selectively examine intrinsic properties. Instead, we have to examine all the intrinsic properties. Call this the total knowledge solution. Now the total knowledge solution seems to help us, because surely if we knew all the intrinsic properties of France, under Von Fintel's broad interpretation of intrinsic, we would know that France is not a monarchy.

But the total knowledge solution fails because now Australia in (14) is in the same boat as France in (1). Total knowledge of intrinsic properties could extend to things like laws, and Australia might have a law that if France has a king in 2006, then they paint all their government buildings blue. But we would also know that all their government buildings are not blue, and no painting projects are under way. Then, by modus tollens, we know that there is no king of France. It could also be that Australia celebrates a national holiday called "End of the French Monarchy Day," that celebrates the end of the French Monarchy. Surely total knowledge of the intrinsic properties of Australia includes what holidays its citizens celebrate.

But perhaps this objection to the total knowledge solution seems strange, because Australia does not have such a law nor celebrate such a holiday. Let me present a clearer example:

(22) The Russian Tsar is on state visit to Ukraine this week.

Now this exactly parallels (14) and hence should be rejected as pragmatically false. But countries in the USSR, before it collapsed of course, celebrated "The Great October Socialist Revolution" or "Revolution Day" (now called "The Day of Accord and Reconciliation" in Russia) on November 7th, which commemorates the 1917 Russia Revolution that disposed the last Russian Tsar. So it is a property of Ukraine that it used to celebrate this holiday. If one were to be a smart-aleck and claim that intrinsic properties only refer to current conditions, then he would be at a loss to explain (i) why past owners are intrinsic properties of a pen and (ii) why this same sentence, uttered in 1960 when Ukraine was part of the USSR and did celebrate the holiday, would be rejected as pragmatically false then but not now.

In lieu of the failure of the total knowledge solution, I put forth my own. I propose that we should simply exclude entities that occur as part of the definite descriptions from being contextually salient entities that we can use to falsify the sentence. That is to say, phrases that function as adjectives on the noun-head of the article "the" cannot be contextually salient entities that can be used to falsify the sentence. Consider "of France" in (1); there "France" is part of a genitive adjectival phrase modifying "the king." Hence France cannot be used as an entity to falsify the sentence. Conversely, "Australia" in (14) and "Ukraine" in (22) can be used for falsification.

Now this solution accounts for all the same intuitions that Von Fintel's does, but it solves the problem of the independence of counter-evidence. Yet it could be objected that the solution is arbitrary and ad hoc. In response to this, I point out that Von Fintel's whole system is constructed as a result of modifying Lasersohn's to account for problems with his own. I also point out that adjectives and adjectival phrases that are part of the definite description are part of what we assume when we assume that the existence presupposition holds. Consider:

(24). The Jewish King of France is bald. Here, part of what we presuppose is not just that there is a king of France, but that he is Jewish. It seems natural that the parts of what compose a non-referring definite description cannot be used to falsify it when we suppose that the whole phrase refers.

I want to flesh out my solution a bit more with regards to relative clauses. Consider:

(25) The King of France, who is eating pancakes in Australia, is bald.

Now (25) mirrors (21)—the Sandy sentence— and should therefore make us squeamish. If it does, then I have nothing to add to my proposal that will interest you. But I, myself, feel that this sentence can be rejected as pragmatically false. I feel that same way about (21). My reasons are, in (25) we could find out that no bald people are eating pancakes in Australia, just as (21) we could find out that Sandy never dates bald men. To accommodate my intuitions, which do not correspond to Von Fintel's, I add an exception to my exclusionary rule so: Contextually salient entities that we can use to falsify the sentence cannot be a part of the definite description, unless they are a part of clause that is embedded within the definite description. I openly acknowledge that this is ad hoc, but it is only necessary to accommodate my own peculiar truth intuition.

My solution allows for Pesetsky's list of the reigning monarchs to stand, because a list that is external to the sentence is obviously not part the definite description. Unfortunately Pesetsky's assertion and Von Fintel's endorsement of it are incorrect. My solution cannot be used in place of Von Fintel's rule that the counter evidence must be such that it could be shown to be true while at the same time the presupposition is not shown to be false. My solution by itself could not account for the following:

(26) The king of France is one of the current reigning monarchs.

Nothing in my solution rules out the set of current reigning monarchs as a contextually salient entity. Yet this sentence is surely one that makes most squeamish. This is because there is no way to show that the king of France is not part of the set of reigning monarch while this showing that the king of France exists. So my proposal is in addition to, not in place of, Von Fintel's other rules for independence of counter evidence.

Finally, there are two ways of reading my exclusionary rule, between which I will remain neutral. Consider:

(27) The king of France lives in France.

If most people reject this sentence as pragmatically false, then my solution will be that entities made salient within the definite description cannot be used as contextually salient entities in the relevant sense, unless they are made salient again, outside of the definite description. If this sentence makes most people squeamish, as I am inclined to think it will, then my solution will be that entities made salient within the definite description cannot be used as contextually salient entities in the relevant sense, regardless of whether they are made salient in another part of the sentence.

V: Summary:

It is clear from the previous discussion of alternatives, e.g. topic focus, Lasersohn's, that Von Fintel's system for explaining our truth-value intuitions is the best one available. Still, it suffers from some flaws, primarily stemming from a poor explanation of the independence of counter evidence and of what counts as a contextually salient entity that can be used to falsify the sentence. Yet, these problems can be remedied by excluding entities made salient within the definite description. Thus his system can still be an effective explanation of our truth-intuitions.

Recognizing Ambiguity
How Lack of Information Scares Us

Mark Clements
Columbia University

I. Abstract

In this paper, I will examine two different approaches to an experimental decision problem posed by Craig Fox and Amos Tversky. This decision problem is intended to illustrate the phenomenon called "ambiguity aversion" that is observed in decision situations under uncertainty. These situations occur when an agent is faced with a decision problem where the probabilities are not specified in advance or readily assessed based on the given information in the decision problem. Because ambiguity aversion in decisions under uncertainty is one of the most fundamental problems of traditional decision theory[1], in attempting to give an account of rational decision making Fox and Tversky offer a hypothesis to account for ambiguity aversion and explain the decision situations that exhibit this problem. After examining their explanation for ambiguity aversion, I will consider another possible approach to these decision problems under uncertainty viz. the recognition heuristic, which has been offered as a decision tool by Daniel Goldstein and Gerd Gigerenzer under an alternate approach to decision theory called "bounded rationality."[2] The purpose of this exercise is to examine how each of these approaches accounts for decision behavior when faced with the same decision problem. In doing so, we will be able to determine how relevant these competing theories are to each other and discover the limitations of the heuristic approach in pursuing a comprehensive model for rational decision making.

II. Framework and Hypothesis

Before proceeding, we need to establish the framework for the hypothesis that is under question here. Fox and Tversky are attempting to identify decision situations where agents exhibit ambiguity aversion. They claim that comparative ignorance effects cause ambiguity aversion and that this is an instance of preference reversal.[3] Fox and Tversky don't attempt to solve the problem of preference reversal; rather, they pose a hypothesis to describe the decision situations where ambiguity aversion manifests. They successfully support their hypothesis in experimental decision situations by restricting ambiguity aversion to situations where comparative ignorance exists. However, this only identifies the

[1] By "traditional decision theory" I mean subjective expected utility maximization.

[2] Simon, 1955

[3] It's debatable if prefrence reversal is confined to comparative ignorance and ambiguity aversion, but this seems to be the approach taken by Fox and Tversky.

domain of relevance for ambiguity aversion while ignoring the more general problem of preference reversal in traditional decision theory.

This reversal of preferences is one of the major problems faced by traditional expected utility theory. It is observed in the Ellsberg problem, where agents systemically violate one of the basic axioms of subjective expected utility theory: Savage's "sure thing principle," or the "independence" axiom. I'm not going to digress into an explanation of this problem because it is a prominent issue in the literature, so prior familiarity with the issues posed by Ellsberg will be assumed here. Essentially, Fox and Tversky claim that the preference reversal is attributed to ambiguity aversion under comparative ignorance.

Because Fox and Tversky fail to address the broader difficulties posed by ambiguity aversion, we might wonder if the recognition heuristic of bounded rationality proposed by Goldstein and Gigerenzer could resolve this problem. If so, how would it resolve the problem? If the heuristic model proposed by Goldstein and Gigerenzer can be applied to Fox and Tversky's decision experiment and develop the same empirical trends[4], then decision patterns that exhibit preference reversal can be accounted for in a theory of rational decision making. The recognition heuristic would provide an alternative method, which would be devoid of the problematic axioms and utility maximization demanded in the traditional theory for explaining such decision problems under uncertainty. I will attempt to apply the recognition heuristic to Fox and Tversky's experimental decision problem to see if it is able to give the same decision outcomes. If this application is successful in this case, then there might be good reason for using this heuristic to supplement the traditional decision theory in situations under uncertainty that illicit irrational choice behavior due to axiom violation. If this attempt is not successful, then the relevance and usefulness of such heuristics in a comprehensive model for decision making must be examined.

III. Fox and Tversky – Ambiguity Aversion

I will begin my analysis by examining Fox and Tversky's experimental setup and the decision problem that is being observed. As stated above, they propose a hypothesis to account for decision situations that exhibit ambiguity aversion. This is their "comparative ignorance hypothesis", which states "ambiguity aversion will be present when subjects evaluate clear and vague prospects jointly, but it will greatly diminish or disappear when they evaluate each prospect in isolation."[5] This is the hypothesis they are testing in their experiment, which presents people with the following decision

[4] That is, the heuristic will also instruct agents to choose the clear bet, which is the decision exhibited in Fox and Tversky's experiment.

[5] Fox and Tversky, 588

problem:

Imagine that there is a bag on the table (Bag A) filled with exactly 50 red poker chips and 50 black poker chips, and a second bag (Bag B) filled with 100 poker chips that are red and black, but you do not know their relative proportion. Suppose that you are offered a ticket to a game that is to be played as follows: First, you are to guess a color (red or black). Next, without looking, you are to draw a poker chip out of one of the bags. If the color that you draw is the same as the one you predicted, then you will win $100; otherwise you win nothing. What is the most you would pay for a ticket to play such a game for each of the bags? ($0-$100)

Bag A	Bag B
50 Red Chips	? Red Chips
50 Black Chips	? Black Chips
100 Total Chips	100 Total Chips

The most I would be willing to pay for a ticket to Bag A (50 red; 50 black) is:___

The most I would be willing to pay for a ticket to Bag B (? red; ? black) is:____[6]

Fox and Tversky conducted this experiment with three groups of people. The first group evaluated the clear and vague bets in a comparative situation as shown above. The other two groups evaluated the bets in a non-comparative situation with one evaluating the clear bet alone, and the other, the vague bet alone. The results of their experiment show that in the comparative group, people priced the clear bet significantly higher than the vague bet; in the non-comparative groups, there wasn't a significant price difference between these two bets.[7] The decision pattern exhibited in this experiment supports their hypothesis that in decision situations of comparative ignorance, ambiguity aversion is exhibited by people preferring the clear bet to the vague bet.[8]

In verifying their hypothesis, Fox and Tversky conclude that comparative situations between one clear option and one vague option exhibit ambiguity aversion, and that this accounts for preference reversal in cases such as Ellsberg's problem.[9] They further conclude that traditional decision theory "requires that the comparative and non-comparative evaluations will coincide" (by virtue of the independence axiom), but it does not "provide a method for reconciling inconsistent preferences."[10] By successfully identifying choice situations of comparative ignorance as exhibiting ambiguity aversion, their results are beneficial to traditional decision theory because they restrict the domain of relevance of the preference

[6] *Ibid.*

[7] *Ibid.*, 589

[8] Preference in the sense that they were willing to wager more on the clear bet over the vague bet. They go on to conduct several other experiments that show similar results in support of this hypothesis.

[9] *Ibid.*, 600

[10] *Ibid.*

reversal problem to decision situations of this type. Although this identification is valuable for traditional decision theory, it doesn't solve the fundamental problem illustrated by Ellsberg. Could bounded rationality's alternative approach to decision making provide a solution to this problem? Possibly, but we would have to determine which tool in the "adaptive toolbox" is appropriate to use in this situation. Now that the decision problem of Fox and Tversky has been identified I will define the recognition heuristic as described by Goldstein and Gigerenzer, provide a justification for attempting to apply it here, and then see if it is, in fact, applicable in this case.

IV. The Recognition Heuristic

In the book Simple Heuristics that Make Us Smart, Daniel Goldstein and Gerd Gigerenzer outline a simple heuristic that describes decision patterns under a specific domain of relevance. To see if we can apply this heuristic to the decision problem presented by Fox and Tversky, it is important to understand how the heuristic works and determine the necessary conditions for implementation.

Since the heuristic functions based on the concept of 'recognition,' this is the first concept that they define. They claim that the term "recognition" has been used in many contexts, but for the purposes of this heuristic, the term is based on the simple binary relationship between the novel and the previously experienced.[11] Simply put, "recognized objects" are objects we've experienced before and "unrecognized objects" are novel objects. With this understanding of 'recognition,' Goldstein and Gigerenzer explore the mechanics of the heuristic by asking an individual to consider the decision problem of "inferring which of two objects has a higher value on some criterion."[12] In evaluating this decision problem, the recognition heuristic says that "if one of two objects is recognized and the other is not, then infer that the recognized object has the higher value."[13]

Goldstein and Gigerenzer demonstrate the use of their heuristic in an experiment involving German and American students deciding which particular American city has the higher population. In their example, the two objects are San Diego and San Antonio, and the criterion is population. The question is then posed: which city, San Diego or San Antonio, has the higher population? According to their results, the German students were as accurate, if not more so, than the American students at deciding which of the two American cities had the larger population. These results are interesting because the German students were more successful in reaching a correct conclusion. The idea is that they were able to use the recognition heuristic

[11] Goldstein and Gigerenzer, 38

[12] *Ibid.*, 41

[13] *Ibid.*

more often than the American students because many of them recognized San Diego, but not San Antonio. From this, the authors describe what they call the "less-is-more effect" which states that more information doesn't necessarily create accuracy and, in fact, there is a certain domain of "ignorance" over which the heuristic is most effective.[14] I won't go into detail about this effect because examining if the Fox and Tversky decision problem satisfies the conditions necessary to implement the recognition heuristic is the only thing I am concerned with. Only after determining if the recognition heuristic can be applied to the Fox and Tversky decision problem will such discussion become relevant.

V. Conditions For Using the Recognition Heuristic

In the Goldstein and Girgerenzer decision problem, the recognition heuristic seems to work rather well. However, it is important to recognize a pair of underlying conditions that must be met in order to implement this particular heuristic. First, the heuristic can only be applied in binary decision situations and second, only in cases where one of the two objects is not recognized. Of course, in the decision problem posed by Goldstein and Gigerenzer, these conditions were satisfied because there was a binary decision between two cities and (in the case of the German students who were able to use the heuristic) one was recognized and the other was not.

The second condition, which I will call the "correlation condition," is that the recognition heuristic is "domain-specific in that it only works in environments where recognition is correlated with the criterion."[15] However, this correlation is not always readily apparent; since the criterion is inaccessible to the agent making the decision, there must be a mediator that exists in the "known environment" in order to correlate the unknown criterion (in this case, population) with the recognized object (San Diego). The mediator establishes this correlation by "having the dual property of reflecting (but not revealing [directly]) the criterion and also being accessible to the senses."[16] In the case presented here, the mediator is the newspaper. There are three variables that describe mediator's relationship between the criterion and the agent faced with the decision problem (this relationship is drawn out in fig. 1 above). These variables are the surrogate correlation, the ecological correlation, and the recognition validity. The surrogate correlation is between the mediator (which is a surrogate for the inaccessible criterion) and the recognized object. In this case, the correlation is between recognizing San Diego and the number of times the newspaper mentions it. The ecological correlation describes the relationship between the mediator and

[14] *Ibid.*, 45

[15] *Ibid.*, p.41.

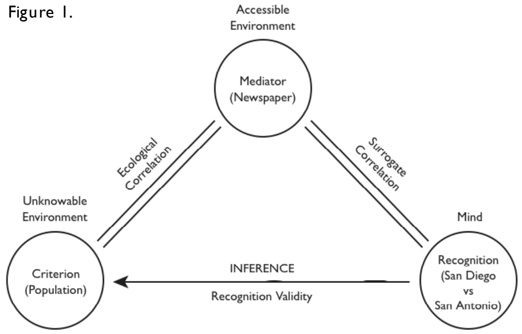
[16] *Ibid.*

the criterion. Here, the population of San Diego is correlated with the number of times it appears in the newspaper without revealing the actual population. Lastly, the recognition validity is the proportion of "correct" answers given by use of the recognition heuristic.[17] Even though the particular decision problem presented here fits the conditions for applying the recognition heuristic, this correlation condition is more complex than the first condition (of binary choice and one recognized object), and so great care must be taken in applying this heuristic to Fox and Tversky's decision problem.

VI. Justification For Applying the Heuristic

Before looking at how this heuristic might be used in the decision problem of Fox and Tversky, it might be appropriate to consider why this particular heuristic seems appropriate to their problem. It is important to provide some justification so that my analysis amounts to more than the random application of a heuristic to an unrelated decision problem. The justification for this attempt is found in the underlying conditions involved with the Fox and Tversky decision problem. Their problem presents an agent with the choice between two objects, Bag A and Bag B, with one choice being clear due to the knowledge of probabilities and the other choice being vague as a result of unknown probabilities. Thus, the Fox and Tversky decision problem seems close to the same types of decision problems that are compatible with the recognition heuristic, viz. a binary choice between one recognized and one unrecognized object. Without any further analysis, one might think that the recognition heuristic is applicable to the Fox and Tversky decision problem, but Goldstein and Gigerenzer state that the recognition heuristic is not a general-purpose strategy, for the necessary correlations do not hold in all domains. However, given the similarity in the decision problems described above, I think this exercise is relevant in attempting to apply this heuristic to the Fox and Tversky decision problem in order to determine how useful or adaptable such heuristics are to similar decision situations. Only by attempting experiments such as these will we be able to know the scope of the domains of relevance for these heuristics. This analysis is especially relevant because if a heuristic as simple as the recognition heuristic is unable to be applied to simple decision situations like the one presented

Figure 1.



[17] *Ibid.*, 42

by Fox and Tversky, (that is, if the domain of relevance is rather narrow) then the practical logistics of the self-proclaimed "fast and frugal" heuristics in the "adaptive toolbox" of bounded rationality will need to be questioned. Before we turn to the heuristics of bounded rationality to solve all of our problems, experiments such as the one attempted here are necessary to determine the limitations of such an approach.

VII. Application Analysis

With this background, it is now appropriate to see if the recognition heuristic can be applied to the decision problem presented by Fox and Tversky. The motivation here is to see if the recognition heuristic can be applied to describe the same decision pattern in order to account for ambiguity aversion in a theory of rational decision making that doesn't run into such problems as preference reversal. This analysis will be attempted by determining if the Fox and Tversky decision problem fits both necessary conditions for implementing the recognition heuristic and then discussing the consequences of the results. Recall the choice problem presented by Fox and Tversky: agents are to decide how much they are willing to pay for a ticket to gamble on drawing a red or a black chip from two different bags. I will assume that the "willingness to pay" as reflected in the price is indicative of which bag the agent thinks is more "likely" to produce a winning result. Thus, a higher price given to one bag over the other reflects the agent's belief that that bag is more likely to win for them than the other. I will use this idea of the "likelihood of winning" as our criterion in the heuristic model. Hence, the decision problem becomes: which of the two bags (Bag A or Bag B) do you think is more likely to produce a winning bet?

In the Fox and Tversky experiment, one group of people chose in a comparative situation, while the other two were in non-comparative situations. Since the recognition heuristic can only be applied in binary choice decisions, it is not applicable to the latter two groups that decided in isolation. I will only be able to look at the decision problem in the comparative group, for that's the only relevant group for this analysis (since I'm not trying to account for ambiguity aversion). I am looking to see if the recognition heuristic will instruct people to choose the clear bet over the vague bet given the same decision problem.

Now, the agent is deciding between two options, "Bag A" where the bet is "clear" because the probability (of drawing a particular color chip) is known to be ½ and "Bag B" where the bet is "vague" as a result of unknown probability. The first condition of applying the recognition heuristic is that the decision problem must be binary and that one of the two objects must be recognized while the other is not. Clearly, the agent is deciding between two objects, "Bag A" and "Bag B", so the binary requirement holds. For the second part of this conidition, we will assume that the agent recognizes "Bag A" as the clear bet by virtue of the known probability, while the agent doesn't recognize "Bag B" as being a clear bet because the probabilities are unknown. Under this assumption, the first condition of applying the heuristic met. Satisfying the correlation condition relies on finding a suitable mediator to correlate the bags with the criterion (the likelihood of winning). In this case, the only possible mediator is the experimental setup because it completely dominates the agent's "known environment" relative to this decision problem. The last step in determining if this decision problem satisfies the conditions for applying the recognition heuristic is if the mediator can establish surrogate and ecological correlations with the recognized object and the criterion (this decision problem is outlined in fig. 2 above). In Goldstein and Gigerenzer's example, they used the frequency of newspaper articles mentioning the recognized object and the frequency of the object appearing in the newspaper to establish these two correlations. The problem here is slightly different. Since there aren't multiple occurrences of "Bag A, clear bet" and "Bag B, vague bet", we can't use frequency to generate these correlations. However, let's assume that this correlation is triggered by the respective probabilities of each Bag, just as the respective frequencies trigger the correlation in the newspaper case. That is, the probabilities presented by the mediator are what triggers the recognition of Bag A being the "clear bet" and Bag B being the "vague bet".

Given these assumptions, is the correlation condition of applying the recognition heuristic satisfied in this case? The surrogate correlation between the mediator and the object relates the recognition of the object to the agent. In the first case, the frequency of newspaper articles about the city was correlated to the number of people recognizing the city. Here, the probability corresponding to the bag is correlated with the number of people recognizing the bag as a "clear bet". Hence, p(Bag A) = ½ , making it the "clear bet", and p(Bag B) = ? making it not recognized as the "clear bet". Though somewhat shaky, this surrogate correlation establishes recognition of the "clear bet" and is provided by the mediator. What about the ecological correlation? This is where the application runs into problems. The mediator needs to be able to establish an ecological correlation by indirectly providing the agent with a correlation between the criterion and the recognized object. In the Goldstein and Gigerenzer example, the newspaper indirectly established the correlation between population (the criterion) and the recognized object (San Diego) by the frequency that San Diego appeared in the newspaper independent of its criterion. In our example, the experimental setup would have to correlate the likelihood of winning with the "clear bet" in such a way that the agent thinks that Likelihood(clear bet Bag) > Likelihood(vague bet Bag). However, there is nothing in the mediator that indirectly (or directly for that matter) implies this because the criterion of Bag A and Bag B are dependant on their respective probabilities and p(Bag B) is not accounted for in the experimental setup.

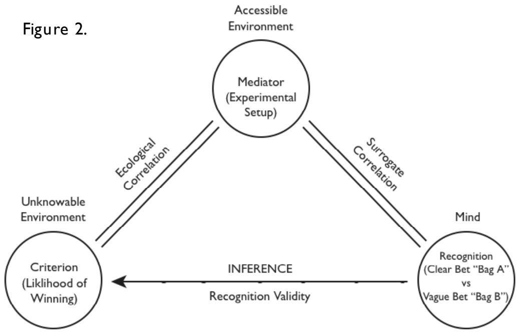


Figure 2.

The failure to establish this ecological correlation makes application of the recognition heuristic impossible in this decision problem. Thus, even though the recognition heuristic seemed to be applicable and satisfied the first condition, it is unable to satisfy both necessary conditions. We must conclude that the recognition heuristic is not applicable to the Fox and Tversky decision problem.

VIII. What Went Wrong?

As stated above, this heuristic only works in a specific domain of relevance. As seen here, Fox and Tversky's decision problem is not in this domain, though the decision problem seemed similar enough at the outset to attempt such an application. One problem in this attempt was that the mediator was unable to establish the ecological correlation, which means that the criterion was not correlated with recognition. This ecological correlation is impossible for two reasons. First, the criterion of the "clear bet" (likelihood of winning) is not independent of the unrecognized object's criterion since the relative probabilities of each bag can determine the likelihood of winning a bet on either Bag A or Bag B. In the Goldstein and Girgerenzer example, it would be like claiming the population of San Diego depends on the population of San Antonio and vice versa. Second, even if we disregard this dependence, the mediator failed to establish an ecological correlation because it didn't provide the agent with the p(Bag B) in order to correlate likelihood of winning with the recognized bag. Of course, if this were the case, then the condition of recognition is also moot and the problem doesn't satisfy the first criterion either. Since the Fox and Tversky choice problem's "knowable environment" was confined to the experimental setup, there is nothing else that can act as mediator here. I think it's important to realize this dependant relationship between the correlation and the mediator because it significantly restricts the domain of relevance for the heuristic.

A more fundamental problem here has to do with limitations on how robust this heuristic is. There are two aspects that limit the domain of relevancy; the definition of "recognition" and the simplicity of correlation. It appears that the definition of "recognition" is more restrictive than the authors let on in distinguishing between the novel and previously experienced. Recognition for the heuristic is merely recognizing an object A based on prior encounters with A. In our example, the recognition assumption was applied to a specific property of the bags viz. being a "clear bet" or a "vague bet," rather than the bags themselves. This property of the bag must be information about the object, whereas recognition only applies to the object itself. "Ignorance" in the notion of "ignorance making us smart" is restricted to instances of recognition. However, ignorance must be more than just not recognizing something, as illustrated in this experiment; it also includes situations where information is lacking. It's apparent that we cannot equate knowledge, or lack of information, with the restricted notion of causal recognition and non-recognition. This concept of recognition used in the heuristic seems too limiting even when attempting to apply it in this simple case.

The second fundamental problem of robustness here is that the recognition heuristic relies on a single recognized object A and one particular attribute X of that object. Even though the Fox and Tversky decision problem was simple and binary, it involves more than the simple correlated relationship between an object A and a specific attribute X. It also involves the interaction of two related objects that are dependant on one another, while the recognition heuristic assumes that the objects under consideration and their criterion are completely independent of each other. Thus, it appears that the domain of relevance of the recognition heuristic is more restricted than was anticipated, as illustrated by its complete failure to give an accounting for the decision problem in Fox and Tversky.

IX. Conclusion

This heuristic is one of the many tools that have been developed in the decision theory known as bounded rationality. The reason that this approach to decision making is interesting is that it doesn't rely on the traditional method of maximizing expected utility Therefore, it's immune to problems such as preference reversal. However, before we view this approach as a panacea for the traditional theory, we must closely analyze these heuristics to determine how relevant they are in certain situations. Determining the relevance of every heuristic out there is beyond the scope of this paper. However, as in the case presented here, we might have good reason to attempt just such a thing. Although, if simple heuristics such as the recognition heuristic fail to apply easily to other simple situations, then their domains are rather restricted. Though the authors admit this, it brings into doubt the general usefulness of this particular heuristic in the so-called "adaptive toolbox".

At this point the question then becomes, 'is the adaptive toolbox offered as a supplement to traditional expected utility theory, or a replacement of it?' If one wants to claim the latter, then the case presented here is a problematic example, because in order give a complete theory of decision making with the "adaptive toolbox" model, there seems to be a logistical problem in implementing and choosing which heuristics have domains relevant to different choice situations. To adopt the imagery of the "adaptive toolbox", it's like needing to carry around an infinite number of screwdrivers for infinite different sized screws. Eventually the toolbox gets so big that it becomes too cumbersome for practical use. We may then need to revert back to the "hammer" of tradition expected utility theory and bang everything out, even if it may not be the ideal tool for the job. Hence, the heuristic approach needs to be supplementary to traditional expected utility theory. Hopefully different heuristics can be developed that can account for situations where the traditional approach fails. As is clearly the case here, the recognition heuristic isn't a satisfactory approach to explain decision patterns exhibiting preference reversal and ambiguity aversion in the context of Fox and Tversky's experiment; so another method must be sought after.

Works Cited

Fox, Craig R., and Amos Tversky. "Ambiguity Aversion and Comparative Ignorance." *The Quarterly Journal of Economics* 1220, no. 3 (Aug., 1995): 585-603.

Gigerenzer, Gerd, and Daniel Goldstein. "The Recognition Heuristic: How Ignorance Makes Us Smart." *Simple Heuristics That Make Us Smart.* Edited by Gerd Gigerenzer, Peter M. Todd, and The ABC Research Group. New York: Oxford University Press, 1999.

Simon, H.A. "Rational Choice and the Structure of Environments," *Psychology Review* 63 (1956): 129-138.

Propaedeutic to a Deduction of Desire in the *Phenomenology of Spirit*

In the opening sections of his chapter on self-consciousness in the *Phenomenology of Spirit*, Hegel describes how consciousness, having superseded the antithesis of the Understanding, becomes self-aware. A crucial component of this moment of consciousness is the appearance of desire, the dialectical necessity of which Frederick Neuhouser attempts to demonstrate in his essay, "Deducing Desire and Recognition in the *Phenomenology of Spirit*." Chiefly because of Hegel's own methodological constraints, the importance of an adequate justification or deduction of the appearance of desire in the *Phenomenology* is very great. In accordance with his rejection of immediate knowledge, it is of primary importance to Hegel's project that the *Phenomenology* explain the development of consciousness without any reliance upon "outside" factors: if any stage of consciousness could be shown to depend upon consciousness' immediate grasp of an outside reality, Hegel's project would fail on its own terms by acknowledging a truth independent of consciousness. As Neuhouser notes in the introduction to his essay, if desire cannot be deduced, Hegel faces the objection that "...its [desire's] introduction into the dialectic is not adequately grounded and that, consequently, Hegel is deceiving us—and himself as well—with his claim to be doing a rigorous and presuppositionless phenomenology" (Neuhouser 243). This objection is especially worrisome considering that desire shows up at the cusp of one of the most important transitions of the entire *Phenomenology*: the emergence of self-consciousness. This transition immediately precipitates the appearance of the lord and the bondsman, Hegel's account of which is probably the most influential ten pages of the entire *Phenomenology,* so a problem with desire could potentially affect our understanding of a good deal more than his own work. Since it is desire that supposedly propels self-consciousness to seek the recognition that will eventually elevate it to the status of Spirit, if desire cannot be adequately explained, neither can the struggle for recognition.

It is in an attempt to steer Hegel clear of such problems that Neuhouser offers a deduction of desire in his essay. After encountering difficulties in attempting the deduction by means of forward movements of the dialectic, he devises and follows a "transcendental" deductive method, concluding that the appearance of desire is justified because desire is a condition of the possibility of the appearance of self-consciousness. In this essay I will show that, while it is not incorrect that there is a necessary connection for Hegel between desire and self-consciousness, Neuhouser's transcendental deduction actually fares no better than his original forward-moving attempt because of a problematic understanding of the nature of desire for Hegel. I will argue that we must think self-consciousness and desire as immediately implying one another, equivalent expressions of a single dialectical moment, rather than as distinct moments separated by a dialectical development as Neuhouser's argument implicitly supposes. We cannot, as Neuhouser suggests, "…accept the initial standpoint of abstract self-consciousness because Hegel has presumably led us to that point along a path of rigor and necessity" (Neuhouser 251), and then proceed—whether forward or backward—to desire, as this will at best only restate the equivalence of self-consciousness and desire. In short, I intend to show that a deduction of desire cannot take self-consciousness as a starting-point, but that it is precisely a deduction *of self-consciousness* that is necessary to justify desire. Finally, at the close of the essay I will briefly outline some of the features that a good deduction will need to have, taking all the above conclusions into account.

Neuhouser begins his investigation with the question, "How is it that the capacity for desire is suddenly attributed to a self-consciousness which at its inception was nothing more than the certainty that it ‘was?’” (Neuhouser 243). With the goal of answering this question, (and having rejected a Kojèvean assertion of desire as a basic and unavoidable assumption), Neuhouser first seeks to derive desire from a forward motion of the dialectic, to discover what about the previous stage of the development of consciousness necessitates the appearance of desire. Failing to find a suitable solution by this method, from which he is able to derive explanations of the role to be played by desire but not the necessity of its appearance, he turns to a different approach. Rather than forward, he proposes to look *backward*, to “…investigate instead the *conditions of possibility* of…” self-consciousness in order to derive desire (Neuhouser 248). Following this transcendental method, Neuhouser proposes to take the emergence of self-consciousness as given and to see whether he can show that desire is a necessary condition thereof. If this can be done, he argues, it will establish the necessity of desire’s appearance. Furthermore, this argument has no need of establishing “…that the point from which it departs is an undeniable facet of our experience” (Neuhouser 251), because to do so would be to embark upon a discussion of the entire *Phenomenology* up until this point, which is unnecessary when only an internal transition is in question.

He goes on to delineate two different transcendental arguments to be considered as candidates for a deduction of desire. First, he suggests that

> We might argue, for example, that the attempt to know can be subsumed under the more general category of “human activity.” Next, we might try to establish that all such human activities are goal-oriented and then that the notion of goal is incoherent without presupposing some sort of desire on the part of the acting subject…The attempt to know is possible only for a being which can desire to know (Neuhouser 249).

Neuhouser rejects this mode of argumentation as unfounded in the text, though he considers it a “not implausible” solution to the problem of desire. On both Neuhouser’s reading and the one I will offer below, this argument does not present itself as a possible solution and perhaps even contradicts what Hegel *does* say, so it will not be considered here.

The second option that Neuhouser considers—and eventually accepts—is to investigate whether desire is a necessary condition of self-consciousness. Here we are to begin from the presupposition of abstract self-consciousness (again, because the entire *Phenomenology* need not be considered for this internal transition).

> Now we ask, however, about the conditions which make possible such a configuration of self-consciousness, and we are shown that such a self-consciousness must also be characterized by desire, for without the structure of desire, consciousness could never have the experience of the other which is a necessary condition for forming a concept of itself (Neuhouser 251).

This differs from the first transcendental strategy because it works with a presupposition having to do merely with self-consciousness rather than with thinking beings as such. The question is not whether we could have even embarked upon the journey of the *Phenomenology* at all without presupposing desire implicitly, but simply whether we can make sense of the appearance of self-consciousness without it. This second option has the clear strength over the first that it needs to presuppose less. It seems much more prudent to simply presuppose self-consciousness than to try to read desire into the entire *Phenomenology* so far, and take it, roughly as Kojève does, as an unavoidable postulate; this second version seems to require less in order to reach the same conclusion. It is also, in a way, consistent with the conclusion I will make below that desire is so closely linked to self-consciousness that the latter cannot be presupposed without the former. Neuhouser proceeds to explain that the contradiction of the moment of self-consciousness (between its identity and its non-identity, or between itself and the living world) is one that must be discovered by self-consciousness, and that self-consciousness cannot make this discovery if it does not first have the faculty of desire, for

> If viewed outside the context of how self-consciousness comes to find this out about itself, there is nothing which one could adduce to argue against this conception of the self. Furthermore, since there is, strictly speaking, nothing "false" about it, it is difficult to see how a contradiction could arise within this mode of self-consciousness which would stimulate another dialectical development of self-consciousness' view of itself. It is only "we" who, having worked through the movement of consciousness, can be shown that this view is inadequate (Neuhouser 250).

Under this interpretation, the ability of self-consciousness to differentiate itself from the living outside world is dependent upon its capacity to desire. The contradiction of self-consciousness is one that emerges only after it has interacted with a world that does not accommodate its desires. According to this view, there is not even an *implicit* contradiction in self-consciousness before this point, but only one which is to arise out of an experience of self-consciousness of which the reader has been foretold. Furthermore, desire is what leads self-consciousness to act in the world, allowing it to establish its self-identity against the outside world:

> It is only when I carry out the same procedure on an experimental level—when I try to make the animal really "for me" by attempting to consume it—that I first encounter the otherness of the animal. It is from its threatening snarl, its attempts to flee—its resistance in general—that I learn that it is something other than myself (Neuhouser 250).

Self-consciousness cannot form without otherness, but it cannot recognize otherness until it asks something of the world, that is, until it desires and acts according to that desire. It must *have* desire before it can interact with the world in such a way that its own separate identity becomes apparent to it, and thus before it can truly be said to have become self-consciousness. Self-consciousness presupposes desire, Neuhouser explains, because otherwise "…consciousness could never have the experience of the other which is a necessary condition for forming a concept of itself" (Neuhouser 251). He then concludes:

> By showing desire to be a necessary condition of self-consciousness, we have in a quite strong sense "deduced" desire. There is nothing arbitrary about its introduction into the dialectic; it is, rather, implicitly presupposed by all that precedes it (Neuhouser 251).

The thrust of this argument is that, because we begin from the assumption that the entire dialectic up to and including the emergence of self-consciousness is necessary, if desire can be shown to be a necessary condition of self-consciousness, it can be said to be necessary to the dialectic as well.

This line of reasoning, however, works only so long as the necessity of desire is not among the presuppositions of which we are assuming the necessity to begin with—and, therefore, only so long as self-consciousness is *not* a necessary condition of desire. If it turns out, not only that desire is a necessary condition of self-consciousness, but that self-consciousness is also a necessary condition of desire, then the assumption of self-consciousness begs the question here. For, in this case, when we assume the necessity of the emergence of self-consciousness and therefore the necessity of all its preconditions, and then "deduce" desire as one of these, we will effectively be restating one of our assumptions. That is, the deduction will have *assumed* the necessity of desire from the start, rendering it just as problematic as the forward-moving deduction that Neuhouser first attempts. Even if its conclusion turns out to be the right one, it will simply be too trivial to be considered a meaningful deduction of desire.

I will now try to show that Neuhouser's transcendental deduction falls into precisely this difficulty. As the problem ultimately stems from a misreading of desire, I will first offer an explanation and critique of Neuhouser's reading. I will then give an alternative one, followed by a sketch of a plan for a forward-moving deduction that seems more promising.

Neuhouser is not entirely clear as to how intimate the connection between desire and self-consciousness is. In one passage he writes that life "…is what intervenes between the stance of merely abstract self-consciousness and self-consciousness as desire" (Neuhouser 246), and shortly thereafter he makes the following point: "Desire appears, then, at that point of the dialectic where the falseness of self-consciousness consists in the fact that it sees its opposition to life without recognizing its essential connection to it" (Neuhouser 247-8). These statements could be read as implying a distance of some sort between a basic form of self-consciousness (as it is immediately constituted in the transition from Understanding) and desiring self-consciousness. In this case, Neuhouser would theoretically allow for a stage of self-consciousness which has not yet begun to desire, by distinguishing an “abstract” self-consciousness from a desiring one. However, he also asserts that

> The relationship is not a simple linear progression in which [naïve self-consciousness] somehow leads to life which in turn becomes desire. Rather, self-consciousness and life are two distinct phenomena which appear on the scene simultaneously, with desire being the moment that mediates between them. The justification for the introduction of desire can therefore not be that it somehow develops *out of* life, as our strategy of looking for rigor within a forward movement of the dialectic would imply (Neuhouser 248).

This formulation is quite explicit that self-consciousness, life, and desire appear all at once, and on this account, self-consciousness never appears *without* desire. But it is unclear why a strategy based upon forward movement necessitates the interpretation, which Neuhouser rightly rejects, of life, desire, and self-consciousness as separate and serial. The possibility that the elements of this triad could be deduced all at once by moving forward from what directly precedes them—rather than trying to deduce them from one another, which, as Neuhouser recognizes, is bound to fail—is not dealt with.[1] Considering the above passage, it is also unclear why Neuhouser even attempts the forward-moving deduction to begin with, since he appears to recognize here that, without a linear progression to work with, it cannot be done. For a deduction of desire from self-consciousness to have any hope of proving the phenomenological necessity of desire, there would have to be some substantial *movement* between desire and self-consciousness; such a deduction presupposes that the connection between desire and self-consciousness is not simply an equivalence, which holds *at a single moment* and therefore has nothing to do with dialectical necessity, but rather that these are two are *ultimately distinct moments*. Neuhouser thus appears to be espousing two opposed interpretations of desire and self-consciousness at different points in his essay: the supposition of *separation* in the dialectical process, which his initial attempt at a forward-moving deduction of desire from self-consciousness presupposes, contradicts his explicit interpretation in the passage above of desire and self-consciousness as inextricably

[1] I argue for this excluded possibility below.

linked. It is this confusion as to the relationship between desire and self-consciousness, rather than some insusceptibility of desire to deduction by forward movement, that dooms his attempt at a forward-moving deduction. Although he seems at one point to recognize the problematic reading of desire underlying his attempt at a forward deduction, he goes on simply to abandon the entire method of forward deduction without recognizing that a backward one could be susceptible to the same problem. Even backward deduction, as I have already shown, simply begs the question when applied to moments between which there is no movement.

Neuhouser's deduction, however, might still be plausible if it could be shown that there is some kind of dialectical distance between self-consciousness and desire. In this case we might argue on Neuhouser's behalf that, while he perhaps has no need to abandon forward movement, his transcendental account nevertheless suffices as a deduction. However, the opposite interpretation—that which has desire as essentially equivalent to self-consciousness—arises far more naturally from the text. There is little evidence to suggest that desire and self-consciousness are distinct in any substantive way, or that we should postulate some sort of distance between them. Instead, they are presented as two descriptions of the very same moment of consciousness, connected by analytical or definitional rather than dialectical necessity. Hegel does not use the language of conditionality with regard to the relationship between desire and self-consciousness; he does not say that one *depends* upon the other for its appearance, but he does repeatedly say that desire is *equivalent* to self-consciousness. How he arrives at this conclusion merits attention.

Hegel writes, in reference to the transition from consciousness to self-consciousness, that "…now there has arisen what did not emerge in these previous relationships, viz. a certainty which is identical with its truth; for the certainty is to itself its own object, and consciousness is to itself the truth" (Hegel §166). Self-consciousness is characterized by the equivalence of the certainty and the truth of consciousness, or the coincidence of certainty and truth in a single entity: consciousness itself. The truth of consciousness has, up until this point, rested in objects exterior to consciousness. Consciousness is both self-certain and the truth of itself, not only for the phenomenological observer, but for consciousness; "…being-*in-itself* and being-*for-an-other* are one and the same" (Hegel §166). The 'I' is the relation of the in-itself and the for-itself of consciousness; it is consciousness' recognition of itself and thus its existence *for* itself. Consciousness says 'I' only when it has first taken the form of self-consciousness: "Opposed to an other, the 'I' is its own self, and at the same time it overarches this other which, for the 'I', is equally only the 'I' itself" (Hegel §166).

But the objects of the "outside" world, with the supersession of Understanding, have also changed their character, and this change has had an effect on consciousness. These objects have not only lost the "simple self-subsistent existence" that they previously had, but consciousness' realization of this loss has brought it back to itself, and "…in point of fact self-consciousness is the reflection out of the being of the world of sense and perception, and is essentially the return from *otherness*" (Hegel §167). That is, self-consciousness is consciousness which looks outward from itself at itself in otherness, but recognizes that self as its own self. It might be described as the ability to think itself in the same manner in which it thinks other objects, and thus the ability to differentiate itself from among such objects and at least the capacity to think what it does as distinct from what they do. It still observes objects of the "outside world" as unities governed by laws in accordance with the preserved moments of Perception and Understanding, but now takes them only as appearances.

Here Hegel introduces desire, as the character of the relation between self-consciousness and its objects which arises out of the need of self-consciousness to establish its essential unity:

> This antithesis of its appearance and its truth has, however, for its essence only the truth, viz. the unity of self-consciousness with itself; this unity must become essential to self-consciousness, i.e. self-consciousness is *Desire* in general (Hegel §167).

Self-consciousness is conscious of two sorts of objects, those of sense-certainty and its own self, but it is conscious of the latter only in opposition to the former, and thus the identity of itself with itself is thrown into question; its objects are both part of it and independent of it, which is a contradiction. The movement of self-consciousness, Hegel tells us, is to be the process whereby "…the identity of itself with itself becomes explicit for it" (Hegel §167). The antithesis of "its appearance and its truth" is the antithesis of self-consciousness and the living world (only now recognized as living), and as an antithesis it must be overcome. It must reconcile its *identity with the world* with its identity as separate from that world, that is, its *non-identity with the world*. Thus what Hegel calls the "…unity of itself in its otherness" (Hegel §177) could also be thought, with Beiser, as "the identity of identity and non-identity" (Beiser 183).

The difference between implicit and explicit self-identity is the difference between the initially unsatisfied self-consciousness and the satisfied one that is eventually to emerge through the process of mutual recognition and ultimately the independence of the bondsman.[2] Here we are concerned only with self-consciousness in its implicit stage, however, as this is the earliest stage which manifests desire:

> To the extent, then, that consciousness is independent, so too is its object, but only *implicitly*. Self-consciousness which is simply *for itself* and directly characterizes its object as a negative element, or is primarily *desire*, will therefore, on the contrary, learn through experience that the object is independent (Hegel §168).

Hegel here indicates that self-consciousness has desire even in its most implicit and abstract form. Desire is a characteristic of self-consciousness before it is explicitly aware from experience

[2] Here I follow Pippin, who writes that "…when he [Hegel] says that self-consciousness 'is only in being recognized,' he means a self-consciousness that is 'in and for itself,' or a finally realized, completed, or reassured self-consciousness. Again, 'self-consciousness achieves its *satisfaction* only in another self-consciousness'" (Pippin 69).

that its objects are independent. From the very moment consciousness becomes self-conscious, even as it is just beginning its quest to overcome its independent object, it is “primarily desire.”

Because of this independence of the object from self-consciousness, self-consciousness posits this object as life, something which contains its own principle of action rather than being a mere adjunct of consciousness. This has the effect that

> …self-consciousness is thus certain of itself only by superseding this other that presents itself to self-consciousness as an independent life; self-consciousness is Desire. Certain of the nothingness of this other, it explicitly affirms that this nothingness is *for it* the truth of the other; it destroys the independent object and thereby gives itself the certainty of itself as a *true* certainty, a certainty which has become explicit for self-consciousness itself *in an objective manner* (Hegel §174).

Hegel reiterates that self-consciousness is equivalent to desire. The two are strongly linked because abstract self-consciousness is by nature the antithesis of itself and the living object, an antithesis which must be overcome just as every other that has occurred so far in the *Phenomenology*. Self-consciousness *is* this antithesis before it *realizes* that it is; it therefore also desires before this realization, for desire is merely the need to overcome this antithesis.

This last point, that the need to overcome this antithesis *is* desire, is essential, because desire is therefore part and parcel of self-consciousness itself. Hegel is not characterizing desire here as either a new “faculty” of self-consciousness or as a phenomenological precondition thereof. Rather, desire is the name given to the necessity of superseding the antithesis of this particular moment in the development of consciousness, an antithesis which is itself the definition of self-consciousness. To return to a passage cited above, “…this unity must become essential to self-consciousness, i.e. self-consciousness is *Desire* in general” (Hegel §167). Desire is merely the expression of the need to overcome the antithesis of self-consciousness and its immediate living objects. Self-consciousness by no means chooses what it shall desire; rather, its desire *is* the necessity of overcoming its inherent contradiction. Desire is this necessity *manifest* through the actions of self-consciousness, and these actions are determined entirely by this necessity. The degree of free agency implied by the idea that self-consciousness has been brought to a point where it is suddenly capable of directing itself according to desires is entirely insupportable at this stage of consciousness, and the text of “Self-Consciousness” gives us no reason to think that this is Hegel’s argument. Self-consciousness may not realize that it has desire before it has become acquainted with the living world, but this does not mean that *that which self-consciousness will realize is its desire* is not present within itself from the point of its emergence. The necessity of overcoming the present moment’s antithesis is as strong (or Hegel means it to be) in this case as it is for every preceding moment, for this preserves the strong necessity of the dialectic itself; we only now call this necessity desire because it is only now recognized as a need. This is not to say that self-consciousness is explicitly aware of its antithesis, that it knows that through desire-directed actions it shall overcome something within itself that is contradictory. Self-consciousness only feels a blind need to do certain things, but *what* it feels this need to do is determined by what only *we* recognize as its contradiction; we can easily say that self-consciousness has desire without saying that self-consciousness knows why it has it. Indeed, we *cannot* attribute such thorough self-knowledge to self-consciousness since this implies a level of free agency in self-consciousness that Hegel has not demonstrated, some criterion other than the one with which self-consciousness is supposed to be working. None of the movements of the *Phenomenology* can be said to be made “intentionally” by consciousness. Such intentional dialectical progression is doubly insupportable because, first, the degree of free agency it implies is not justified, and second, it waters down what is supposed to be the very strong necessity of the dialectic by making it contingent on what consciousness “wants.” We can only conclude that the opposite is true: what consciousness wants is necessitated by its constitution, not by something that it learns once it has already become self-consciousness. Self-consciousness does not *phenomenologically* condition desire; rather, they are two interchangeable descriptions of the same moment. Desire is the name given to the need of self-consciousness to overcome the untenable antithesis of self and outer living world, the antithesis which simply *is* the definition of self-consciousness. Thus we should not ask why self-consciousness desires to supersede its objects, for the need to supersede these objects, the mere fact of being self-conscious, is itself desire.

We can easily adduce good reasons why this need to overcome an antithesis is only now called desire. Only now is this need of consciousness—the necessity of overcoming antithesis, the same necessity which has been in force all along in the *Phenomenology* so far—something of which consciousness is aware; the need to supersede the antithesis is now a need not merely "in itself" but a need for self-consciousness. Desire is thus a dialectical necessity like any other except that it is for self-consciousness, which is the same as to say that self-consciousness is for self-consciousness. Action arises with desire, for, as Beiser writes, "The ego now has to begin acting since action is the decisive test for its thesis" (Beiser 180) that it is all reality. Indeed, "action" in the strict sense is impossible before the emergence of self-consciousness because in this moment consciousness first realizes that it is an entity distinct from the world in which it operates, and that a certain *mediation* is necessary between what is willed and what is actual; self-consciousness just *is* the conceptual separation of willing and actuality in which context alone the idea of action makes any sense.

We can conclude overall that Hegel's repeated statements to the effect of "self-consciousness is Desire" must be taken literally to mean that desire and self-consciousness are inseparable, two aspects of one and the same moment of the dialectic. To call a thing desiring is to call it self-conscious *and* vice versa. As I have shown, even in its basic, abstract and unsatisfied phase self-consciousness is characterized by the antithesis of itself and its immediate object (life), because this antithesis is the definition of self-consciousness. Precisely the need to overcome this antithesis, I have argued, Hegel calls desire. It is *not* that self-consciousness is aware of the antithesis *as such* and desires to overcome it; rather, the desire of self-consciousness is simply the expression of this necessity, which abstract self-consciousness does not explicitly comprehend. Self-consciousness in general thus immediately implies desire. However, because desire is differentiated from the need to overcome any previous antithesis by the very fact that it is a need which is only now *for consciousness*—which is recognized *as a need*—desire cannot occur outside of the self-consciousness which experiences it as its desire; desire also immediately implies self-consciousness. Self-consciousness and desire can thus be considered to mutually condition one another; as Kojève writes, "…the self-conscious being, therefore, *implies* and *presupposes* Desire" (Kojève 4, emphasis mine).

This reading, if correct, rules out the first interpretation of desire that appears in Neuhouser's essay (as distinct from self-consciousness) in favor of the second (as an equivalent formulation of self-consciousness). To be sure, the biconditionality of desire and self-consciousness for which I have argued here is in accord with Neuhouser's claim that we cannot try to deduce the necessity of desire in a simple forward progression *beginning from self-consciousness*; on both his account and mine self-consciousness implies desire. But if my account is accurate, we also cannot deduce desire by a backward-looking movement from self-consciousness, for on my account self-consciousness also *presupposes* desire. That is, I have aimed to refute the assumption underlying Neuhouser's attempts to move *between* self-consciousness and desire—whether by forward- or backward-looking analyses—that these two elements are somehow distinct in the dialectical process. A deduction of desire *from* self-consciousness would require that the former be a dialectical consequence of the latter rather than the analytical equivalence that (I have argued) Hegel's text gives us. In other words, the text makes a tautology of Neuhouser's deduction. His early attempt at a forward-moving deduction naturally does not prove the necessity of the appearance of desire in the *Phenomenology*, for his initial assumption of self-consciousness, because of its analytical reducibility to desire, is also merely an assumption of desire. However, if I have been correct here, a similar problem belies his transcendental account as well.

But the question now becomes: how do we deduce desire, given that it both conditions and is conditioned by self-consciousness even in that moment's most basic and "abstract" form? The clearest solution is that, rather than assume the transition from Understanding to self-consciousness as given and trying to deduce desire from self-consciousness, which will give us only a trivial result, we should look at the transition from Understanding to examine *its* necessity. If, as I have argued, self-consciousness immediately implies desire without further dialectical movement, then we will be unable to accomplish any deduction of desire's necessity by assuming self-consciousness as given. In order to make a meaningful and philosophically interesting deduction we must rather demonstrate the necessity of the entire movement from Understanding to the moment of self-consciousness and desire. Although such a deduction is a separate issue from what I have undertaken here, it is not difficult to delineate how it could proceed. The necessity of the movement from Sense-certainty through Perception to Understanding could be assumed. This is, of course, only a general guideline, and as I hope the present discussion has made clear, great care would need to be taken as to where *precisely* the assumption of necessity should stop and the deduction should begin. Certainly it would begin no later than the very first appearance of the form of Understanding, or perhaps even the movement of Perception which segues into this appearance, so as to avoid the risk of inadvertently taking as given any development that needs to be proven. From this point, one would need to demonstrate that the movement from Understanding necessarily results in the form of consciousness that Hegel describes in the closing paragraphs of the section on Understanding and in the subsequent discussion of self-consciousness. Such an investigation would thus be focused primarily on the text of "Force and the Understanding" and perhaps the close of "Perception" and the opening of "Self-consciousness." If the conclusions of this discussion have been correct, such an investigation would have no need for recourse to Neuhouser's transcendental method (though it may be valid). One would also need to be careful, as with any deduction of necessity in the *Phenomenology*, not to take anything as a "given."[3] This includes refraining from attributing motives or faculties to consciousness for which Hegel has not provided clear justification. Though it is difficult to talk about the machinations of consciousness without resorting to the language of independent agency, it would need to be kept in mind that words like "tries," "struggles," and (this especially in the case of a deduction of desire) "wants" are only imperfect terms that reflect our inability to escape the point of view of the phenomenological observer; their usual implication of the kind of agency we like to attribute to ourselves should by no means be taken up along with them. Finally (and again if I have been right to argue this), a deduction of self-consciousness from a framework such as this one itself constitutes a phenomenological deduction of desire. Once the moment of self-consciousness has been deduced, no further dialectical movement is necessary in order to reach desire.

I have not myself attempted to deduce desire or anything else in this discussion. I do hope, however, to have clarified how such a deduction would have to be accomplished, by

[3] Neuhouser rightly finds Kojève's characterization of Desire (as an "undeducible" but necessary element of the *Phenomenology*) to be far from Hegel's intention and thus unacceptable for the purposes of the deduction he undertakes (Neuhouser 245).

referring to certain problematic elements of Neuhouser's own. I have argued that Neuhouser's reading of desire fails to properly recognize the strength of the relation between self-consciousness and desire. I have proposed a reading of Hegel that has desire as an integral part of self-consciousness, one which outwardly manifests the inner contradiction in self-consciousness even from its initial emergence. Desire and self-consciousness are thus not separated by any dialectical development but are two aspects or formulations of a single moment. Under this reading of Hegel, Neuhouser's original attempt at a forward deduction and the transcendental or backward-looking deduction he eventually rests upon fail for the same reason: the equivalence of self-consciousness and desire makes a tautology of any deduction of desire that takes self-consciousness as its starting-point, whether it proceeds forward or backward. Finally, and in light of these conclusions, I have offered a basic overview of how a deduction of desire is to be successfully carried out. Rather than deducing, or even beginning to deduce, the necessity of desire in the *Phenomenology*, I hope this essay succeeds at providing something like a propaedeutic to a deduction of desire. Such a deduction is of particular importance to a good understanding of Hegel's work, but it begins where this discussion leaves off.

Works Cited

Beiser, Frederick. 2005. *Hegel*. New York: Routledge.

Hegel, G. W. F. 1977. *Phenomenology of Spirit*. Trans. A. V. Miller. New York: Oxford.

Kojève, Alexandre. 1980. *Introduction to the Reading of Hegel*. Ed. Allan Bloom. Trans. James H. Nichols, Jr. Ithaca: Cornell University Press.

Neuhouser, Frederick. 1986. "Deducing Desire and Recognition in the *Phenomenology of Spirit*." *Journal of the History of Philosophy* 24: 243-62.

Pippin, Robert. 1993. "You Can't Get There From Here." In *The Cambridge Companion to Hegel*. New York: Cambridge. 52-85.

Does Anomalous Monism Have Explanatory Force?

Andrew Wong
Washington University, St. Louis

The aim of this paper is to support Donald Davidson's Anomalous Monism[1] as an account of law-governed mental causation in a world unfettered by psychophysical laws. To this end, I will attempt to answer one principal objection to the theory: the claim that Anomalous Monism lacks sufficient "explanatory force."[2] Though not quite the standard objection, I believe it to be the most formidable, and hence the most crucial to address.[3] The argument's strength is that it need not dispute Davidson's assumptions. It accepts Anomalous Monism as an internally consistent theory and attempts to show that what follows is an account in which mental events routinely cause actions, but can never manage to explain them. If the objection is right, Davidson's theory has clearly fallen short of explaining mental causation in a satisfactory way. The success of Anomalous Monism, then, requires the falsity of the explanatory force objection. I argue that a proper construal of Davidson's principle of rationality will show the objection to be misguided.

> "Mental Events" reconciles the paradox which arises from three principles Davidson held *ex hypothesi*: (1) Mental events interact causally with physical events (*Principle of Causal Interaction*), (2) Where there is causality, there must be a law (*Principle of the Nomological Character of Causality*), and (3) There are no strict deterministic laws on the basis of which mental events can be predicted and explained (*Principle of the Anomalism of the Mental*).[4]

Tension is apparent in that some mental events must interact causally with physical events and thereby feature in laws, and that this is an explicit contradiction of the Principle of the Anomalism of the Mental. Accepting three further principles, in addition to the three above, will resolve the tension: (4) Each mental event is token identical with some physical event, (5) Causality (and identity) relations hold between individual events no matter how described, and (6) Events instantiate laws – and can be explained or predicted in light of laws – only as described. Thus, the Principle of Causal Interaction requires causal participation of events regardless of mental or physical description, the Principle of the Anomalism of the Mental pertains only to events described as mental, and the Principle of the Nomological

[1] All references to Davidson's original presentation of Anomalous Monism are from Donald Davidson, "Mental Events," reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers (Oxford: Oxford University Press, 2002).

[2] To my knowledge, one of the strongest formulations of the "explanatory force" objection comes from Louise Antony. I will be addressing her arguments more or less directly, and taking them to be representative of objections of this type. The reader is therefore urged to see Louise Antony, "Anomalous Monism and the Problem of Explanatory Force," The Philosophical Review, Vol. 98, No. 2. (Apr., 1989), pp. 153-187.

[3] The standard objections to Anomalous Mo

[3] The standard objections to Anomalous Monism, which are the various epiphenomenal accusations, are close relatives to the explanatory force objection. The main difference, as I see it, is that an account may successfully establish the causal efficacy of mental events without establishing the way in which mental events rationally explain actions. This is essentially Antony's argument. Though my paper is not meant to directly address the epiphenomenalist objections, my argument will, mutatis mutandis, apply to many formulations of it. This will be clear enough to anyone familiar with those objections if and when it should occur.

[4] Davidson, *Mental Events*, 116

Character of Causality requires that two events in a causal relationship have *some descriptions* which instantiate a law.[5] The tension has dissipated, for now an event with a mental description (i.e. a mental event) may interact causally with physical events without violating any principles, so long as the mental event has a physical description (i.e. is also a physical event) which features in a strict law.

The explanatory force objection, which argues on the basis of the above characterization, is divided into two parts. In the first part, Antony argues that speaking of causally efficacious physical events in psychological does not explain the resultant event in all cases. Despite the truthfulness and rationalizing power of the mental descriptions of causal events, effects are only sometimes thereby explained.[6]

But what are rationalization and explanation? A rationalization simply refers to an instantiation of the Principle of Causal Interaction; it suggests a causal connection between a mental event and a physical event. By the Principle of the Nomological Character of Causality, any such connection implies the presence of a law. The important thing to note is that folk psychological (i.e. "commonsense") rationalizations may obey such laws – and hence be causal – despite the universal absence of laws framed *in* mental terms. An explanation, on the other hand, is different. Not every true causal statement is a causal explanation. Explanations are intensional, which is to say that the cause of an event is an explanation of that event if and only if it is picked out in a particular way. [7]Antony's example[8] showcases this intentional aspect of explanation while introducing us to the first part of her objection:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally.[9]

Antony proposes that in cases like these, where the causal chain linking reasons to actions is somewhat amiss, rationalizations will not work as explanations. After all, Davidson only had two conditions for an adequate rationalization:[10] the principle of rationality and the causal condition. The principle of rationality says that "we cannot intelligibly attribute any propositional attitude to an

[5] *Ibid.*, 119

[6] Antony, Anomalous Monism and the Problem of Explanatory Force, 183

[7] Antony, Anomalous Monism and the Problem of Explanatory Force, 163-164

[8] Antony actually uses two examples in her papaer. The second example deals with a hurricane causing a disaster, and with the reporting of each event in the newspaper. Perhaps the second example shows some kind of problem, but I do not see how any problem with non-mental events illustrates anything at all about mental events. In any case, the exclusion should not weaken any arguments. As Antony notes, both are used to show the same problem (*Anomalous Monism and the Problem of Explanatory Force*, 168).

[9] Davidson as quoted in Antony, Anomalous Monism and the Problem of Explanatory Force, 167

Antony finds three conditions in Davidson's "Actions, Reasons, and Causes." I see no problem with her three criteria. As far as I can tell, mine and hers are equivalent, since Davidson's principle of rationality would have us view another person as "a believer of truths" (Mental Events 123), which entails her first principle (viz., that the attributed mental attitudes be true). I formulate the requirements in the way I do after the fashion of "Mental Events." Nothing of consequence to the argument will be lost if either criteria are substituted.

[10] Antony finds three conditions in Davidson's "Actions, Reasons, and Causes." I see no problem with her three criteria. As far as I can tell, mine and hers are equivalent, since Davidson's principle of rationality would have us view another person as "a believer of truths" (*Mental Events* 123), which entails her first principle (viz., that the attributed mental attitudes be true). I formulate the requirements in the way I do after the fashion of "Mental Events." Nothing of consequence to the argument will be lost if either criteria are substituted.

agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions… (t)he content of a propositional attitude derives from its place in the pattern."[11] The causal condition says that "the event cited as the reason in the explanans is the cause of the event cited as the action in the explanandum."[12] Even though both conditions are met here (the belief-desire pair attributed to him does not make him incoherent, and the belief-desire pair caused the action), citing the belief-desire pair alone would plainly be inadequate, for that is not the whole story about *why* he dropped his companion. The reasons cause the action in the "wrong way."

What Anomalous Monism needs but lacks, Antony contends, is an account of how reasons and rationalizations come together as the causes of actions, or how "reasons can have causal efficacy *in virtue of their reasonableness*."[13] As things currently stand, it seems that the logicality of the mental description of a physical event dictates the causal connections into which it might enter, which would be strange, indeed. Antony wants to know what, precisely, the relationship is between the physical descriptions of one's body, inside and out, and the "commonsense" explanations of one's behavior.

> In order to drive the problem into even further clarity, Antony asks us to:imagine another climber, a vicious one this time, who, having the same desire to be free of her partner, *deliberately* does what she believes will fulfill that desire, viz., lets go of the rope. The problem for Davidson here is to say why rationalization is a proper explanation in the second case but not the first.[14]

Antony uses these counterexamples as a lead-in to a "property theory" of psychological properties which she thinks is needed if Anomalous Monism is to hold up against these examples. If I can show, however, that Anomalous Monism is able to address these examples with its current machinery, Davidson would have had to commit to no such theory.

I submit that the stories of the two climbers are incomplete, so that it is no wonder there seems to be something lacking. In addition to the belief-desire pair in the thought experiments, we must mention *all other relevant beliefs and desires* for each climber – and we should certainly hope to find that they are multitudinous and diverse. For if the two climbers had all of the *same* beliefs and desires, including the two mentioned (as hypothesized by the examples), then in what sense is one "vicious" and the other merely "unnerved"? For that matter, how can one act deliberately, and the other unwittingly?[15] Each climber's rationality necessitates a divergent set of propositional attitudes. To the one we must allow the afflicting fear of guilt over another's demise, and to the other we rightly impute an unrivaled egotism. We need not accept these specifically, but only in some such contexts would their actions make sense.

If the unnerved climber had no other desires or beliefs with which the particular belief-desire pair in question could hang

[11] Davidson, Mental Events, 122

[12] Antony, *Anomalous Monism and the Problem of Explanatory Force*, 166

[13] *Ibid. 168*

[14] *Ibid.* 173

[15] Of course, it may well be the case that both climbers are equally vicious, and that the unnerved one was merely denied the opportunity to exhibit his truly nefarious character. But I doubt this is what either Davidson or Antony mean by the examples. Nearly needless to say, I assume that the first, unnerved climber would not have intentionally dropped his companion, even had he been free from his overwrought event.

upon, becoming "unnerved" would be truly inexplicable. Observe that such a reaction follows only if he holds something like the belief that dropping his companion would be terrible. A climber who believes that dropping his companion is *in his companion's best interest* would have no reason to become unnerved upon the realization that such an outcome was within his power. If this is at all accurate, then the original belief-desire pair can *only* cause the action when other relevant beliefs and desires are present, and once those other beliefs and desires are present, the action will be explained. This is just what Davidson's principle of rationality said from the start:

> There is no assigning beliefs to a person one by one on the basis of his verbal behavior, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest.[16]

The reason, once we know its full mental context, does not cause the action in a "wrong way" at all. The examples, as used, begged the question against Anomalous Monism, for they did not properly acknowledge the principle of rationality. When we recognize this principle, all mental causes of an agent's action *must* be sensible causes. It follows that if the action *can* be explained by the attitudes of the agent, it *will* be so explained. If it is not explained, then we must have failed to construe the agent's actions in the most cogent way. The failure of the example was simply that it did not take into consideration enough of the agent's attitudes. Reasons have causal efficacy "in virtue of their reasonableness" when and only when they are globally consistent with the agent's rationality.

I have attempted to show that Davidson's principle of rationality is not only useful in establishing the "indeterminacy of translation," but that it aids us in difficulties like the one above.[17] I will now show that it can help us to answer Antony's other objection, too – one which *does* concern indeterminacy.

In a Davidsonian model, if one reason R_1 (and not R_2) causes Hermione's action, it is because R_1 is identical with some physical event *c*, and *c* causes the agent's action.[18] But a key part of the principle of rationality is that it prohibits us from settling on any one interpretation of an agent's propositional attitudes:

> when we use the concepts of belief, desire and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory… a right arbitrary choice of a translation manual would be of a manual acceptable in light of all the possible evidence, and this is a choice we cannot make.[19]

Because of this indeterminacy, Antony argues that the question of which propositional attitude is identical with the neurophysiological event *c* will be answered differently every time, and hence there are no *real* psychoneural identities – not even token ones. The lack of "genuine facts" about an agent's mental events dictates a lack of "genuine facts" about their relations with physical events. If we attempt

[16] Davidson, *Mental Events*, 122

[17] All I mean by this is that the principle of rationality is not only useful for establishing that there are no strict psychophysical laws. See Davidson, Mental Events, 123

[18] Antony, *Anomalous Monism and the Problem of Explanatory Force*, 183

[19] Davidson, Mental Events, 123

to hypothesize a fact of the matter that *c* is identical with some particular mental event, we will have settled on some theory of the agent's rationality, no doubt without the necessary full evidence. In light of all this, an Anomalous Monist cannot rely on causal links to rationalize actions – and hence cannot account for explanatory force – since there are no objective identities.[20]

How, then, can we simultaneously harbor the freedom, rationality, and efficacy of the person? I want to suggest that Antony has confused epistemological indeterminacy with metaphysical indeterminacy. While our *theory* of interpreting Hermione's propositional attitudes may be "radically indeterminate," it does not follow from this that Hermione's propositional attitudes *are themselves* radically indeterminate. Davidson's principle of rationality was an instruction to "third persons" to be charitable when *assigning* attitudes to agents, not a statement about the natures of agents' psychologies. Thus, it is not necessarily the case that there are no genuine facts about the contents of Hermione's mental events. In fact, we *must* assume that Hermione's attitudes and actions form a coherent pattern if we are trying to mirror that pattern with an evolving theory. If Hermione's mental events were themselves indeterminate, then no amount of evidence could give us reason to favor one translation manual over another.

If we assume a fact of the matter about Hermione's mental events while maintaining the radical indeterminacy of all theories *about* her mental events, the other difficulties quickly evaporate. Hermione has a mental event. Hermione performs a physical action. That physical action must have a cause. The cause must be physical. Hence, the mental event is (token) identical with the physical cause of the physical action. But here's the kicker: *Because* any interpretation of Hermione's mental events will still be radically indeterminate, all speculation about the token identities will be indeterminate as well. The token identities are unknowable but present, and that's precisely what Davidson said all along.

[20] Antony, *Anomalous Monism and the Problem of Explanatory Force*, 183-184

Works Cited

Antony, Louise. "Anomalous Monism and the problem of Ezplanatory Force," The Philosophical Review, Vol. 98,No. 2. (April, 1989), pp. 153-187

Davidson, Donald, "Mental Events," Philosophy of Mind: Classical and Contemporary Readings (edited by David J. Chalmers). Oxford: Oxford University Press, 2002

Moral Luck and the Function of Results in Punishment[1]

Keith C. Hemmert
Harvard University

Most people believe that a person can be held responsible only for what is within his or her control. A person cannot be held accountable – legally or morally – for events over which he exerted no influence. This intuition is known as the Control Principle. However, the strict application of the Control Principle seems to diminish the scope of human agency to a vanishing point, and thus eliminate the possibility of moral responsibility. Cases of moral luck occur when events of luck or chance to play a substantial role in moral evaluation.

Put another way, whether or not someone is good or bad depend (at least in a large part) what that person does – not on what happens to that person. If we allow events of luck (things that the agent didn't control) to play a role in moral assessment, then the moral assessment of that agent depends, at least in part, on luck. But it seems odd to think that whether or not someone is blameworthy or praiseworthy is just a matter of luck.

Overview of the problem

The Control Principle Cases of moral luck are particularly troubling in part because of the intuitively plausible idea that people cannot be morally culpable for events that they did not cause, or for events caused by factors beyond their control. It is easy to see why this idea is so appealing.

Moral evaluation does not stand independent of agents. When we exact a moral verdict, we are not judging a set of circumstances in the absence of an agent. The presence and actions of an agent are integral to moral evaluation. [2] Without a rational agent, we do not have an appropriate forum for moral evaluation; the actions of an agent are tied to moral assessment. The agent must have in some way caused (or played a role in causing) the thing that is subject to moral evaluation. We do not consider the lottery winner to be morally praiseworthy (insofar as he won the lottery), and we do not consider the innocent bystander at a car wreck to be morally blameworthy (insofar as he did not cause the wreck). In order to deserve praise or blame, the agent must have had control over the events in question.

This does not amount to a defense of or argument for the control principle, but rather an explication of its intuitive appeal. Indeed, it is difficult to imagine world in which we the ascription of moral responsibility did not require a causal connection between an agent and the actions or circumstances of moral evaluation; without such a restriction, any agent could be held responsible for any action or event. [3] Ann, sitting quietly reading, could justly be blamed for the baseball flying through the window of the library – as opposed to the

[1] This research was conducted with the generous support of the Edmond J. Safra Foundation for the Center for Ethics at the John F. Kennedy School of Government, Harvard University in the form of a Lester Kissel Grant in Practical Ethics.

[2] Without claiming whether or not 'willing' can be properly called an action, let us use the word 'action' loosely so as not to exclude acts of willing.

[3] This argument is based on common sense notions of causation. I accept that it is possible for rational agents to be directly responsible for specific actions in the broad sense that our actions are not merely the results of a series of electrical and chemical reactions in our brains, but rather the results of a rational deliberative process.

boys who let their outfield abut the library. Or, to use a less hyperbolic case, imagine holding a doctor responsible for an infection after an operation, despite the fact that he operated flawlessly. The Control Principle is a basic part of our understanding how actions and events happen in the world – individual agents are responsible for causing them.

This intuition is, as I said, intuitively acceptable and appealing. The problem arises when

> the condition of control is consistently applied, it threatens to erode most of the moral assessments we find it natural to make. The things for which people are morally judged are determined in more ways than we at first realize by what is beyond their control. And when the seeming natural requirement of fault or responsibility is applied in light of these facts, it leaves few pre-reflective moral judgments intact. Ultimately, nothing or almost nothing about what a person does seems to be under his control.[4]

As Thomas Nagel eloquently put it, once we examine exactly what it is that we have control over, it starts to look like we don't actually have control over very much. Nagel considers four kinds of cases in which this is true. [5]

Luck

- Resultant luck. Much of how our actions actually wind up influencing the world is beyond our control. Two agents might take exactly the same actions, but each with totally different results. Smith and Jones both drive home intoxicated. They drive equally recklessly, and are equally lacking in motor control. On Smith's route home, a little girl happens to be playing in the street, and he hits her. On Jones' route home, no one darted into the street. The two agents' actions were identical, but the results of their actions were quite different – and, vitally, the difference was caused by factors outside of the agents' control.
- Constitutive luck. Disposition and personality are beyond the influence of our will. While dispositions and inclinations might change over time, constitution in this sense refers to precisely the parts of personality that are beyond the scope of control. We can act kindly, but cannot have a kind personality by sheer force of will; our natural disposition is largely beyond our control and the result of fortune. Yet we are sometimes morally assessed for our disposition despite its being the result of fortune rather than our will.
- Circumstantial luck. To a large degree, the situations we face are beyond our control. Smith's path to work takes him past a lake where, one morning, a child is drowning. He has the opportunity to demonstrate his moral praise- or blame-worthiness. Jones' path to work, however, doesn't pass near the lake, and he is never faced with the same situation.
- Antecedent luck. The previous four sorts of luck push us to hold agents morally responsible for only the pure acts of will; after all, luck plays to great a role to hold them accountable for the results of their actions, or their character, or even their moral transcript. But their will itself may be the product of antecedent causes

[4] *Moral Luck* in Mortal Questions. Thomas Nagel. (New York: Cambridge University Press, 1979). 26.

[5] The following four varieties of luck are proposed by Nagel in *Moral Luck*.

> outside the agent's control. The causal factors can range from concerns about strict determinism to more everyday things, such as who your third grade teacher might have been. Whether the worry is classic concerns of free will or more mundane causal factors, the 'acts of will' that we want to be responsible for could well be things that we have no control over at all.

After considering these four types of luck, it begins to look like we have control over very little of what we do. What makes the problem so challenging (and so interesting), of course, is that it arises from some very common-sense and intuitively acceptable notions. The very thing that leads us to the paradox – the Control Principle – was the thing that seemed at first to be an intuitive and elemental part of moral responsibility.

Facets of the problem

1. Agency reduced to a vanishing point.

Moral evaluation aside for a moment, Nagel's lucid exposition of the problem reveals fundamental problems with our basic understanding of human agency. While he acknowledges the age-old problem of strict determinism, this is not the strength of his point. He makes no attempt to answer that question; instead, and more importantly, he highlights the ways in which more everyday varieties of chance seem to diminish the scope of human agency until it disappears.

This, of course, is a deeply unsettling thought. The problem of determinism is so deeply troubling precisely because it eliminates agency. But determinism has its stalwart opponents, and cogent arguments can be marshaled against it. Nagel's points are so worrisome because they derive from perfectly ordinary, everyday concerns.

An all-encompassing answer to the worry of moral luck will paint a full, compete picture of human agency. Part of that must be a stand on how to understand causal factors. Such a discussion would be outside the scope of this project.

2. Moral concerns

The puzzle of moral luck raises serious concerns about moral judgments. Moral evaluation is an integral part of everyday life. It expedites our interactions with the world; the system of rules and guidelines that morality affords us helps us avoid dealing with every situation on an ad hoc basis. Cases of moral luck pose a serious problem for our moral compass. They highlight what seems to be a paradox in morality: that the rigorous application of an intuitive feature of moral responsibility – control – winds up eliminating the very possibility of assigning moral responsibility.

Consider the following case, which serves to illustrate the problem:

> **The drunk driver**: Consider Smith, who drives home from the pub after having drank quite a lot. He is in no condition to drive. On his trip home, he speeds, weaves all over the road, runs stoplights, and generally exhibits those signs that are consistent with drunk driving. But Smith makes it home without injuring anyone or harming anyone's property. He did endanger quite a few folks, including himself, but no tangible harm resulted from his drunken escapade.

Now imagine a second driver, Jones. Jones leaves the pub in the same condition as Smith; he's had too much to drink, and is in no condition to drive home. He drives home just as dangerously as Smith. We stipulate that the two men are equally drunk (have equally impaired motor functions), equally lacking in control of their vehicles (and that the vehicles are identical), etc. But while Jones is driving home, a little girl happens to be playing in the front yard of her house, and she runs into the road chasing her ball just as Jones is driving by. Being as drunk as he his, he cannot stop – he strikes and kills her.

Both Smith and Jones had the same level of control (or lack thereof) over precisely the same things: their decision to drink and drive, their ability react upon seeing a pedestrian, etc. It was strictly a matter of chance – outside the control of either driver – that there was (or wasn't) a little girl playing at the time he went by.

Yet Smith and Jones will (for the most part) be judged quite differently. Smith will not escape a negative moral evaluation – he endangered his own life and the life of those around him – but he will not be judged as harshly as Jones, whose actions resulted in the loss of a life. But again, the only difference between the situation of the two men – the presence or absence of the little girl – was completely and totally outside of each of their control; it was strictly a matter of luck.

The case presents the difficult question of what constitutes moral reactions (i.e., Jones is worse than Smith because he killed a girl), and what constitutes emotional reactions (i.e., Jones behavior sickens/angers me to a much greater extent than Smith's because it resulted in a loss of life). Perhaps the moral evaluation of the two agents should be identical – after all, they each took the same risks – and it is appropriate that our emotional reactions should differ so much. The problem of differentiation is relevant to understanding what constitutes moral judgment.

I said earlier that the problem of moral luck poses challenges to moral evaluation. But it is not at all clear what moral evaluation is. When we say that, "Smith should be judged more harshly than Jones," what are we talking about? Are we morally evaluating Smith (or Jones), or Smith's actions in this particular case? Are we to understand Smith's poor decision-making and reckless driving as evidence of his character, or do we evaluate those actions independently of character? What is it that we are doing when we blame Smith and Jones, and is it the same thing as moral evaluation? And (this is perhaps the fundamental question of the problem of moral luck), why is it that we sometimes think that responsibility is necessary in order to assign blame? Efforts to answer the above questions about the nature of praise, blame, and moral evaluation will shed light on the role of the Control Principle in moral assessment, and help to resolve the problem of moral luck. I will take up these questions in Chapter 2.

3. Legal problems

While the problems that moral luck poses for morality are daunting, they are not the same as the problems posed to legal theory. While law does reflect morality in a very general way, the law must consider a variety of concerns with which moral theory need not bother (e.g., the practical viability of enforcement). Do questions of moral luck truly influence practical ethical dilemmas in the law? They do, but perhaps not to the same extent, nor in the same manner, that they influence the more theoretical aspects of moral theory. The challenge posed to the law is somewhat less extensive than the challenge posed to common sense morality, but a challenge nonetheless.

The law, while not a direct correlate of morality, is nonetheless generally interested in setting standards of socially acceptable behavior. The law lays down rules of behavior, and works toward specifying the repercussions for those who fail to comply with those rules. It is tempting to look at law as the practical version of morality, or to try to link law and morality in a strict way. But while law and morality undoubtedly have links, they are not one and the same. As Justice Oliver Wendell Holmes put it in Commonwealth vs. Kennedy, "the aim of the law is not to punish sins, but is to prevent certain external results."[6] So methods and concerns of moral judgment do not necessarily transfer to the law.

Law, unlike pure philosophy, does not have the luxury of entertaining serious doubts about human agency. It must assume that individuals do possess free will, and are capable of rational deliberation about action. Given its practical aims, our legal institutions must also draw lines – lines that are usually implicit in the law itself, and in decisions of more difficult cases that highlight ambiguities in the law – about the limits of luck.

The practical aims of the law, though, are far from clear. Different schools of thought advance different aims for the law. Deterrence theorists argue that the law serves to maintain public good, and that punishment is both a personal and general deterrent to bad behavior. Retributive theorists maintain that crimes upset the balance of benefits and burdens in society, and that punishment serves to restore that balance. Others claim that punishment serves a rehabilitative purpose; criminals are punished so that they might learn why their behavior is indeed a moral transgression. [7]

Under different conceptions of the purpose of the law and the purpose of punishment, the cases of moral luck resolve themselves differently. Some legal theorists argue that antecedent luck can and should enter into the law. The argument of these sort of theorists goes something like this: since antecedent causes form one's moral outlook, and since such antecedent causes are beyond an individual's control, the law must consider such causes exculpatory. Those who had the misfortune to be born into poor social situations cannot be held responsible for their skewed moral outlook; that skew was caused by factors beyond their control, namely their rotten social background. Thus, society is more to blame for the individual's moral flaws – society at large is responsible for its failures, such as the ghettos – than the individual himself. (Hence the argument for the role of prisons not as fundamentally punitive institutions, but rather as educational institutions).

While I will not surmise all of the arguments for and against such theories, the 'rotten social background' theory does serve to illustrate why concerns about moral luck do raise substantial challenges for the law. Fundamental questions about the purpose of law and punishment are pertinent to answering the questions that moral luck poses.

The way in which the moral luck problem raises questions about the law is actually quite similar to the way in which it raises questions about morality. Moral luck forces us to reconsider the basic questions about moral assessment: what is it, and why do we engage in it? Efforts to answer these questions will help us take a defensible stand on the moral luck cases. Similarly, the problems that moral

[6] *Understanding Criminal Law*, Third Edition. Joshua Dressler. (New York: Lexis Publishing, 2001). 108.

[7] See, for instance, Jean Hampton, *The Moral Education Theory of Punishment*, in Punishment.

luck poses to the law make us reconsider basic questions about what it is that the law is trying to accomplish. Efforts to answer that question will help us resolve the challenges to the law posed by moral luck.

Resultant luck in particular raises some difficult problems for legal theory. On the one hand, it seems bizarre to punish actions taken in good faith with good intentions, but which result in negative consequences due to factors outside an agent's control. To hold the agent legally responsible smacks of injustice, and is difficult to justify by either deterrence theory or retributive theory.[8] However, those negative consequences nevertheless were the results of the agent's actions, and it seems appropriate that a) we need to deter not only those sorts of consequences, but also that sort of risk-taking, and b) the balance of benefits and burdens has been upset, and must be restored. The problem here is not altogether dissimilar from the moral one.

Resultant luck brings to mind two legal concepts: negligence and strict liability. The concept of negligence features in both tort law and criminal law. Civil negligence is "a deviation from the standard of care that a reasonable person would have observed in the actor's situation."[9] Criminal negligence is "… a gross deviation from the standard of reasonable care… a person is criminally negligent if he takes a substantial, unjustifiable risk…."[10] The concept of negligence involves risk taking; when chance plays a role in determining the outcome of those risks, we have cases that looks quite similar to moral luck. Negligence itself is a controversial topic of the criminal law. Some deterrence theorists argue that precisely because negligence is the failure "to perceive the risks… of conduct," it cannot be deterred. Many retributivists believe that "the basis for just punishment is voluntary wrongdoing," and since negligence lacks a voluntary element, it cannot justifiably be punished.[11] Negligence (and its relevance to moral luck) will be further discussed in Chapter 4. A crime typically has two components: mens rea and actus reas. Actus reas is the 'physical or external portion of the crime,' and mens rea is the 'mental or internal' component.[12] This is just to say that crimes are willful, voluntary acts done intentionally by a rational agent that result in harm. Strict liability doctrine is an exception to this general principle: "a strict liability doctrine is a rule of criminal responsibility that authorizes the conviction of a morally innocent person for violation of an offence, even though the crime, by definition, requires proof of a mens rea."[13] Strict liability doctrine, when applied, allows individuals to be held criminally responsible for results that they did not intend; when those results came about (at least in part) as by chance, then we again have something that resembles a case of moral luck posing a challenge to the law.

There is a body of case history that involves instances in which resultant luck posed challenges to the law. Here I'll briefly describe one of those cases. It is a civil case, but it serves to illustrate how resultant luck can pose challenges to the law. [I hope to replace this with a criminal case, and eventually discuss only the criminal law. In the meantime, though, Palsgraf does illustrate the difficulty of causation, luck, and responsibility in the law.]

Palsgraf v. The Long Island Railroad Co. In Palsgraf, a man carrying a package rushed to board a train as it was departing the platform.

[8] That is not to say that it is impossible to justify. Both deterrence theorists and retributive theorists can marshal arguments to do just this. I discuss and argue against them in Chapter 3.

[9] Dressler, 128

[10] *Ibid.*, 130

[11] *Ibid.*, 128

[12] *Ibid.*, 81

[13] *Ibid.*, 143

The man nearly fell off the train, but a guard reached toward him and pulled him in as a guard on the platform pushed the man from behind. In the course of these actions, the package that the man was holding fell. Its outward appearance gave no clue as to what it was (it was wrapped in paper). It contained fireworks, which exploded when the package fell. The explosion caused scales at the other end of the station to fall, striking and injuring the plaintiff. The plaintiff wanted to hold the guard (and hence the railroad) responsible for her injuries.[14]

Palsgraf is a case of resultant luck. The relevant result (the scales falling and striking the plaintiff) of the guard's actions (trying to help a man board the train) was beyond his control. The court ruled that the guard could not have known what was in the package (nor, of course, could he have known that it would fall, although the risk of the package falling is entailed in helping him onto the train). The case is landmark, as it goes a long way toward explicating causation in the law. The guard's actions, while necessarily a link the chain of events that caused the plaintiff's injuries, were not the proximate cause of those injuries.[15]

As this brief sketch of the problem shows, moral luck poses challenges to a wide range of philosophical and legal questions. It touches on basic problems of free will and abstruse questions of liability in the law. A complete and total solution to the various facets of the problem would solve some of the most difficult philosophical dilemmas that have persisted for centuries. Perhaps most importantly, it would resolve the basic paradox posed by two common sense notions of morality.

Here I wish to give a brief sketch of a possible solution for one very narrow aspect of the problem. As I described above, moral luck poses slightly different problems to legal theory as it does to moral theory. However, those legal problems are nonetheless quite important. I will attempt to raise some possible strategies that could be used to justify different punishments for identical acts with different results (cases of resultant luck).[16]

The Function of Results Punishing for Negligence

I want to be clear at the outset that the problem I am attempting to solve is not the moral one. Rather, I am concerned in this section with problems of resultant luck in the law. Specifically, how might we be justified in meting out different punishment to two agents whose identical actions resulted in quite different outcomes? I will assume as a background general deterrence theory. In addition, I take it to be clear that the state is justified in punishing in order to deter citizens from egregious risk-taking. With this as background, I'll proceed.

From a historical standpoint, results are vitally important to criminal law. As described above, the traditional definition of a crime must involves an actus reas – it is necessary that there be some harm that results from an agent's actions. In cases resultant luck, however, it is mere chance that determines whether or not there is a harmful result (or the degree to which

[14] See <u>Philosophy of Law</u>, 598

[15] Causation in the law is a thorny topic. A body of cases exists which goes a long way toward explaining the principles of causation in the law. It is too great to discuss here, but suffice it to say that those cases do not resolve the problem satisfactorily, else the problem of resultant luck in the law would no longer be challenging.

[16] I will leave out of this discussion any significant treatment of Nagel's main point – the diminishing of human agency to a vanishing point. It goes beyond the scope of this paper, and does not, as I discussed above, play as relevant a role in the legal problems as it does in the moral ones. The law does not have the luxury of taking those doubts overly seriously, lest law cease to an effective means of molding behavior.

the result is harmful).

This definition – that a crime involves a mental state as well as certain physical actions and the relevant results – is central to the defense of different punishments for identical acts. Without both components, it will become necessary to draw a line about what degree of mental activity or physical result should merit a given punishment. However, the placement of such a line will always be arbitrary. The least arbitrary place to draw the line – and hence the most just – is at the differing results of actions. Here I will attempt to show, though a series of cases, why it becomes problematic to punish similar acts identically.

1. Consider Matt and Doug, who decide to get completely drunk at a party.[17] In this intoxicated condition, they find the host's rifle collection, and decide to see who is a better shot by trying to hit the streetlight. Neither hits the streetlight, but one of their bullets – they cannot tell whose – strikes and kills a bystander. Only sophisticated ballistics tests can determine which gun the bullet came from.

It is tempting to argue here that both Matt and Doug should suffer the same punishment. They took the same reckless actions, endangered the same people, and had control over the same factors. It seems as though there is no morally relevant difference between their actions; after all, it might have been either of them who killed the bystander. Even though only one of them is actually responsible for the loss of a life, the sole factor responsible determining which of them did so was pure chance.

2. Now imagine that, rather than both being completely drunk, Matt is slightly more sober than Doug. When they discover the rifles, neither Matt nor Doug has any hesitation about shooting at the streetlight to test their prowess with firearms. But Matt, being slightly more sober, is capable of aiming more accurately. Though it is still impossible to tell without the sophisticated test which gun the bullet came from, it does turn out that it is Doug's gun, not Matt's.

Here there is a relevant difference in their behavior and capabilities, but it is not clear what impact that should have on their punishment. Matt, despite his relative sobriety, was still enthusiastic the reckless activity of shooting at the streetlight. While his aim was better, his judgment was not. Should he suffera lesser punishment?[18]

3. Now imagine that Matt is completely sober, and Doug is completely intoxicated. Upon discovering the rifles, both Matt and Doug think that shooting is a good idea (Matt just has poor judgment). As above, it is Doug's bullet

[17] This example is drawn from Richard Parker's *Blame, Punishment, and the Role of Result* in Philosophy of Law.

[18] This variation on the case raises an interesting consideration: in many aspects, it is impossible to know whether any two individuals were indeed acting identically. In this case, perhaps the police at the scene could have administered BAC tests to determine that Matt was less drunk. But perhaps they didn't – how are we (or, more importantly, the jury) to know that he was less drunk? Perhaps the only evidence that we (or a jury) can have is the very fact that he didn't fire the shot that killed the bystander.

More generally, the argument goes that the result is the only sure way that we can justifiably judge someone to have been so lacking in control that they deserve the utmost punishment. Without the dead body, there is no way to measure just how much the agent was endangering people. Of course, on this view, individuals are not punished for *the results of their actions*, but rather for the endangerment (recklessness, etc.) that their actions caused. The result merely serves as evidence of the level of endangerment (i.e., that it was sufficient to result in harm). This is problematic, as we usually tend to believe that a murderer is punished *for murdering someone*, not for endangering someone to such and such a level that they could have (and did) die. See Norvin Richards, *Luck and Desert*.

that kills the bystander. But Matt, of course, is highly complicit in the course of events unfolding as they did; he not only failed to stop Doug, but also egged him on and participated himself. While Matt and Doug's actions are quite different, Matt is nonetheless a factor in the events that caused the bystander's death – and his behavior is just as morally blameworthy. Should his punishment be the same as Doug's? Probably not, but he is certainly not undeserving of blame, and perhaps even some punishment.

4. Here Matt is sober and Doug is drunk, and upon finding the rifles, Matt is cognizant that trying to hit the streetlight is a bad idea. He discourages Doug from trying. But Doug is a particularly belligerent drunk, and insists that they shoot. Matt storms away in frustration, leaving Doug to his own devices. One of Doug's shots strikes and kills a bystander.

Matt played a minor role here in events that caused the bystander's death, but he remains complicit. He failed to take the appropriate actions to stop Doug; he could have argued with him longer, been more forceful in his discouragement, or physically prevented Doug from taking the rifle.

5. In the last case, Matt takes the most responsible course of action. As Doug tries to get his hands on a rifle, Matt argues him out of it, eventually physically removing him from the room. Doug never gets the chance to demonstrate his poor aim, and the bystander remains unharmed.

The purpose of these cases is to highlight the difficulty and arbitrariness of drawing the line of when to punish two agents similarly. It is clear that in case five that Doug and Matt do not deserve the same punishment (indeed, since no crime has been committed, no one may be punished in case five) – yet nor are they equally blameworthy. But case one tempts us to say that they should be punished identically. Does the change occur in case two? After all, they weren't equally drunk, and so they didn't endanger the bystander to the same degree. But epistemic considerations give us pause (see footnote 16); how do we know that they weren't equally lacking in control? How can we know the difference of degree to which they were endangering others? And how do we know that, in case one, there weren't some differences in the extent of their control? (Alternatively, how can we know that they were identically reckless?)

The difficulty in saying precisely where the shift occurs makes the necessary arbitrariness of drawing the line apparent. However, the result (in this case, knowing which gun the bullet was fired from) eliminates some of that difficulty. That is not to say that drawing the line at results is completely non-arbitrary. But the result certainly removes some of the epistemic difficulties as well as some of the arbitrariness of the problem. Being the least arbitrary, it is the most just. It provides the best basis – that is to say, the best evidence – for a decision on how to punish.

Another way of putting this (one that is intuitively appealing) is to say that, despite the fact that all appearances indicate that Doug and Matt acted identically, only one of they was actually responsible for the death of the bystander. Both caused significant endangerment; only one caused a death. Allowing the result to factor into the different punishment of the two agents satisfies an intuitive urge; it allows the special connection between an agent, his actions, and the results of those actions to remain intact.

I should note that this does not eliminate the controversy around negligence. In the series of cases illustrated above, neither agent had a mens rea – a guilty mind. Certainly their actions were the proximate cause of the bystander's death, and those actions were voluntary, but they did not intend to bring about a death. The series of cases does demonstrate the intuitive appeal of criminal negligence – most would tend to think that drunkenly firing rifles into a populated area constitutes a crime – it does not amount to an argument.

Some closing remarks

Moral luck is a phenomenal problem. It touches on a tremendous range of philosophical problems, from esoteric questions about free will, to more common sense concerns about the scope of human agency, to questions about intuitive moral reactions (and the extent to which we can trust them), to legal questions that have a very real impact on the way our society functions.

Here I have attempted to give a comprehensive overview of all of those facets of the problem. While the questions of agency and the purely theoretical moral questions are not irrelevant, they are nonetheless not as pressing as the challenges posed to our legal system. I take the most significant of those challenges to be the claim that we should punish without significant regard for results. The aim of the law is to mold behavior of the individuals that make a society. Without taking a stand on the retribution versus deterrence debate, I think it is clear that neither pure deterrence nor pure retributivism suffices to defend out current penal institution.

Our current institution places high regard for proportionality of the crime to the punishment – both deterrence theorists and retributive theorists can agree on this. While endangerment without harmful result and endangerment with harmful result are both legally and morally condemnable, they are not the same thing. To ignore that difference is to increase the arbitrariness of our legal system, and to decrease the extent to which we insist that evidence and fact determine punishment. Though it is tempting from a moral standpoint to ignore results and focus on intention and action, to do so would be so impractical as to result in a lesser degree of justice.

On Foolishness: In Arguments We Must Value Only the Truth

Robert Trueman
Fitzwilliam College, Cambridge University

I would like to offer some advice as subversive as when Russell suggested that we believe only what we have reason to. When entering a debate, inquiry, investigation or any other activity involving the conflict of different beliefs, that we do so with an attitude of impartiality until, upon summary of the evidence and the arguments, we find one belief to be superior to another.

Most of us are convinced that whatever we think is right, by virtue of the fact that we think it. After all, if we take everyone's self-assessment to be accurate, none of us are even remotely foolish. Only foolish people entertain false beliefs and, therefore, none of us entertain false beliefs. Perhaps we could lend this belief some credence if it were not for the fact that people hold views which are a flat contradiction to others'.

So where does the falsehood lie in our valid syllogism? To be sure, sadly, it is not the case that only foolish people entertain false beliefs; perhaps foolish people merely entertain more false ones than true. However, I have a suspicion that this is not the only fault. If it were, according to our own assessments, that none of us are foolish, therefore none of us entertain more false beliefs than true, and therefore there ought to be a great deal (although not unanimous) agreement on all subjects. Unfortunately, especially in matters of importance, agreement is rare and where it is found it is rarely whole-hearted; people hold back a small portion reserved for a subtle modification that only they, as such supremely unfoolish people, can make. This all leads me to consider that at least some of our self-assessments are wrong. Some of us are foolish.

Obviously you and I are not foolish, but many others are. Even more unfortunate for them, they believe that they are not foolish and, perhaps do so with the same firmness as you and I. When they consider their beliefs, so obviously false to you and I, they honestly believe them to be plainly true. In fact, when they survey our superior beliefs, these poor, misguided people may even accuse us of being fools.

From these considerations it is plain that we cannot rely on our convictions, and their prima facie reasonableness, as a proof that they are true. We must instead consider all the arguments and evidence available to us, and then not reach our conclusion until new arguments or evidence become available. Surely, this imperative is declared more often than it is followed. To the common man, even the possibility of subjective ethics is so ridiculous that it warrants no consideration. Sadly, the undergraduate is often found to forget difficult questions in opposition to her essay's argument when she produces it. Even some professional academics, Heaven forbid (assuming I may posit such a place), may be reasonably accused of similar crimes.

It may be sensible to ask what has caused this sad state of affairs, where reasonable people preach something that they do not practice. At least one answer, it would appear, is plain: When entering any conflict of beliefs, most of us do so with an agenda. This agenda may be as mild as perpetuating the beliefs we find most appealing. At other times, our motivations are more sinister, like monetary or political profit. When we do this, we are likely to focus our minds (or at least our performance in the conflict) on the strengths of the beliefs that we are supporting and the weaknesses of those that we oppose. We will rarely consider our beliefs' problems or our opponents' advantages; when we do, we rarely travel very far. We think of minor setbacks in our own beliefs, which are easily repaired, and trivial superiorities in our opponents', which are deftly dealt with. Should we ever find devastating flaws in our own beliefs or an unconquerable strength in our opponents', we are as likely to ignore them as we are to actually accept the consequences.

Some might be thinking that although it is true that most of us do suffer from this vice, it is no sad situation. Darwin's theories have been applied to so many other fields, so why not here as well? There are many different beliefs, and sometimes they come into conflict. Fortunately for beliefs, they all have weapons and defences. Their weapons are arguments which show the inadequacies of other beliefs in the same subject, increasing the likelihood that this particular belief is true. They have two defences; arguments which show their strengths and counterarguments to the weapons of other beliefs. When two or more beliefs find themselves in some contest, unless they are evenly matched, one belief normally defeats the other by virtue of their defensive or offensive capabilities. Once a belief has been defeated, one of two things will happen: the belief is left for dead, or it is modified in such a way that it becomes stronger. No matter which is the case, we end-up with a stronger body of beliefs than before. Either the weak mutate or they die. It is through this survival of the fittest that the best beliefs are created and therefore the current state of affairs, insofar as it facilitates this evolution, is a good thing.

And perhaps if these claims were true, the current situation would not be too bad a thing; however, defeated beliefs do not always get left for dead or evolve into something better. Sometimes the empty carcasses of beliefs are hauled about as if they were still alive or they are transformed into confusing, disfigured creatures that only appear more convincing as they have become harder to understand. People refuse to let go of a belief in which they have a vested interest. This often results in people holding onto beliefs with little or no reason, or even onto beliefs which have been conclusively shown to be false. Others inject liberal amounts of sophisms or delicate complexities into their beliefs that create a façade of reasonableness but add nothing to the substance below the surface. Successful beliefs are not necessarily truer than those they defeat. The evolutionary struggle of ideas is not a guarantee of enlightenment.

Having explained why the present situation is a problem, it would seem proper to offer a solution. Mine is a simple one: when entering a conflict of beliefs, one's only agenda should be to find the most-likely-true beliefs. If we were to enter all arguments, debates, and so on without a prejudice for one belief to defeat another, then we could create impartiality. Rather than being trapped in the point of view of one belief, we would move freely between them all. In doing so, we would view and even invent arguments from each side objectively, by considering their merits. After this, we will be able to draw some conclusion until some new evidence is revealed or argument created. As these conclusions would be based on a full consideration of all the relevant factors available to us, they would be more likely to be true than the beliefs we entertain now. This would be the case even if the conclusions we arrive at after this practice are the same as the beliefs we currently hold, as this method would give us more reason to believe that they are true.

It would, of course, be foolish to think that this process can be performed in all conflicts of beliefs. In law, politics or even most every-day arguments, people are not particularly interested in reaching beliefs which are more likely to be true than any others. They are concerned with some gain they can bring themselves, whether it be success, power or merely the ability to say, "I was right." The problem for these people is that perhaps because of the competitive nature of conflicts, they have a sense of winning and losing. They find that they are drawn to one point of view more than any other, for whatever reason, and if that opinion is shown to be mistaken or inferior then they feel as though they have lost. As such, their aim of reaching beliefs which are most-likely-true is lost in the shadow of the desire to win.

For this last group there is, I hope, a solution. Rather than being in the mind-set of having opponents in a contest of pride, they should consider winning to be the completion of the one positive agenda, that is the aim of reaching the most-likely-true beliefs. If this occurred, for example, atheists would not find in believers an enemy to defeat, but instead, a partner in the search for most-likely-true beliefs. Stripped of this competitive element, the atheist and the theist would be able to properly and impartially consider one another's views—a most desirable state of affairs that, for the most part, does not currently exist.

There is one objection which, although troublesome, I must address if I do not wish to be foolish myself. If we should enter each conflict of beliefs with impartiality, eager to consider everyone's views in detail, then should we waste our time on things such as myth and legend? If we say "no", then it would seem the only reason that we do so is because we consider such views to be plainly ridiculous; however, if we are allowed to make such judgements about the prima facie reasonableness of such views, then why not all others?

My answer is that we should, if we are at all concerned with having true beliefs about such things as mythical creatures, give an impartial survey of all the relevant views as I have advocated so far. For instance, the only evidence available to me about the existence of mythical creatures consists of second-hand reports of legends in which they play such a central role. The evidence against their existence is the fact that I know that the surface of the world has been for the most part, although not entirely, explored and even in the places where mythical creatures are reported to have existed no convincing evidence has been found to indicate that they do. It does not take many moments to conclude, impartially, that the evidence available to me is stronger on the side of their non-existence. Should, however, someone present to me more evidence and arguments in favour of mythical creatures roaming the world, then insofar as I am interested in having most-likely-true beliefs in regards to mythical creatures, I should impartially consider the evidence, whether this takes a longer or shorter time. If such inconveniences are the price of having the most-likely-true beliefs in other, perhaps more important, subjects, then I for one will happily pay it.

If more of us followed the humble suggestion of this essay then, perhaps, more of our self-assessments would be accurate, and fewer of us would be foolish.

(Un)Doing Critical Philosophy: Reflections on Adorno's *Aesthetic Theory*

Larry McGrath
University of California, Berkeley

An Introduction to Adorno's Thought and Elucidation of Terminology

Theodor Adorno (1903-1969) was the foremost theorist of the Frankfurt School. Originally established as an institute of social research, the Frankfurt School evolved as an informal grouping of German philosophers and social critics, including Max Horkheimer, Walter Benjamin, and later Herbert Marcuse and Jurgen Habermas. The Frankfurt School advanced a critical theory of society, which deepened theoretical understanding of late capitalism, the rise of European fascism, and the state of late modernity. This proceeded by synthesizing elements of Marxism, Psychoanalysis, and the German critical tradition (inaugurated by Kant). The core of this position was arguably best captured in Adorno and Horkheimer's *Dialectic of Enlightenment* and Horkheimer's *Traditional and Critical Theory.*

For Adorno, critical theory grounded itself as a historically self-reflexive sociological critique. As a mode of philosophical-sociological inquiry, critical theory takes aim at the ideological, scientific, and economic conditions of post-industrial capitalism. These conditions reflect the development of refined instruments of social control that prolong and intensify mass unfreedom and absorb avenues of resistance, such as those originally articulated by Marx.

My interpretation of Adorno's thought thus foregrounds its Marxist elements. These elements coalesce in the task of critical theory: to imagine the world differently by subjecting it to critical negation. As I argue in this essay, *Aesthetic Theory* advances an understanding of art against the backdrop of a world in which this task becomes increasingly difficult to fulfill. *Aesthetic Theory* thus integrates the principles of critical theory with an account of art in its contemporary context. What is art? How does art affect us in the modern era? And, most significantly, how should philosophy engage art? These questions guide *Aesthetic Theory* in its effort to revive the power of philosophy and art in an era that blunts the critical potential of both.

1. Decline of metaphysics ("a post-metaphysical world"): refers to an ontology of becoming, which this paper credits to Friedrich Nietzsche's assault on metaphysical certainty. This view holds that there exists no abstract (Platonic) realm of stability beyond the physical world; instead, the world is composed of struggle among competing forces. Thus, physics best describes our world, as one in constant motion, reducible only to the activity of which it is composed.

2. Instrumental rationality: can be understood as a mode of thought that converts the inherent value of something into its use-value. Money, for example, is purely instrumental; its only value is its ability to purchase other items, as no bill is worth anything beside the instrumental use to which it is put. For Adorno, instrumentalization is the dominant mode of thought in late modernity.

3. Violence of the concept: refers to the consumption of all objects under the

conceptual control of the subject. This is an epistemological counterpart to instrumental rationality, which imposes the concept of identity[1] on things to render them know-able. We come to know society and nature through identity-thinking when all that is real must harmonize with our own conceptual system.

4. Negative dialectics: attempts to limit the dominance of identity-thinking by reflecting on the limits of our concepts: to think what is un-thought. The task is to negate the identity of what appears immediate to us, and therefore think of what falls outside identity. For Adorno, negative dialectics is the central task and driving thrust of critical theory.

The beginning moments of *Aesthetic Theory* offer the author's reflections on the contemporary status of art; "nothing concerning art is self-evident anymore, not its inner life, nor its relation to the world, not even its right to exist."[2] My project takes Adorno's reflections as its point of departure. I will begin by identifying the socio-historical conditions that efface art's self-evidence: a post-metaphysical world circumscribed by instrumental rationality. It is in response to these two conditions that Adorno crafts his negatively dialectical aesthetics. The question I am interested in asking is how philosophy responds to these same conditions. In order to answer this question I probe the style of the philosophical text *Aesthetic Theory*. My analysis will uncover a negatively dialectical style of philosophical construction that adapts itself to the philosophic conditions of a post-metaphysical world and resists instrumentalization by the demands of instrumental reason. This will unfold as I trace the development of the central theses of *Aesthetic Theory*, analyzing the style of the text, which opposes the traditional logical and narrative form of philosophical texts.

I. Historically philosophy has built theories of aesthetics upon an ambivalent relationship between their constituents, philosophy and art. On the one hand, art, in its production and reception, is often understood to be inherently subjective. The subject's aesthetic experience of its engagement with an artwork seems to be confined to the particularity of that experience. On the other hand, philosophy aims to conceptualize what is universal in that experience. Adorno recognizes "the fundamental difficulty, indeed impossibility, of gaining general access to art by means of a system of philosophic categories." At the same time "aesthetic statements have traditionally presupposed theories of knowledge."[3] Thus, aesthetics must negotiate these twin dimensions, the philosophic demand to articulate universal categories and the particularity of the work of art. This duality motivates a dialectical aesthetic that mediates between the philosophical concept and the work's resistance to conceptual consumption.

Adorno's dialectic mediation between these oppositional dimensions owes its groundwork to the seminal aesthetics of Kant and Hegel. Kant's contribution to the aesthetic tradition is his transcendental critique of aesthetic judgment. In the *Critique of Judgment* judgment is a timeless faculty of knowledge; it functions as

[1] Identity is that which renders an entity definable and recognizable. This is an axiom of logic, according to which the identity relation holds only between a thing and itself: x = x.
[2] Aesthetic 1

[3] *Aesthetic* 332.

"the ability to think the particular as contained under the universal."[4] Aesthetic judgment, in particular, grounds itself in the a priori faculty of taste, confirmed by "the fact that whenever we judge any beauty at all we seek the standard for it a priori in ourselves, and that the aesthetic power of judgment itself legislates concerning the judgment as to whether something is beautiful or not."[5] Thus, the subject who deems a work of art beautiful does so on the basis of universal and necessary conditions. Hegel's' philosophy of art responds to Kant's transhistorical account of aesthetic judgment. For Hegel, art has a history of its own. It's understanding is not guided by universal categories; instead, Spirit speaks through individual works within their historical era. In its opposition to the Kantian primacy of the subject, whereby the concept subsumes the particular work of art, Hegel's aesthetics reveals the works own cognitive comportment.

Drawing upon these contributions, Adorno crafts an aesthetics that builds upon the autonomous historicity of the work of art, yet recognizes that the subject who engages the work cannot dispose of its conceptual apparatus. The work is a product of history, but comprehension of its uniqueness depends upon its subjective mediation. *Aesthetic Theory* orients this subjective mediation in new directions attuned to the intrinsic temporality of the work. He writes of aesthetics, "as an investigative procedure subsumption never reveals aesthetic content, but if subsumption is rejected altogether, no content would be thinkable."[6] Hence, what is necessary is an aesthetics that curbs the conceptual subsumption of the work. Dialectics, in its movement between the work and the concept, responds to this demand to preserve the particularity of the work in the face of the violence done to it by the concept: "Aesthetics is not obliged, as under the spell of its object, to exorcise concepts. Rather, its responsibility is to free concepts from their exteriority to the particular object and to bring them within the work."[7]

This is a daunting project that Adorno takes up. The trick is to deploy philosophy successfully against its own medium : the concept. But before moving to whether Adorno succeeds in his task, we should note the socio-historical circumstances that an aesthetics must also address. These are the historical conditions philosophy now faces in a post-metaphysical world and the barbarity imposed on that world by the logics of late capitalism. In the face of such conditions, aesthetics finds its task to "free concepts from their exteriority" more demanding.

The decline of metaphysics marks the rise of a world wherein the stable ground upon which to found an aesthetics dissolves. Aesthetics can no longer ground itself in the lofty Kantian position of a transcendental subject. This is because faculties of knowledge do not submit themselves to transhistorical investigation. Nor can Adorno work within the logics of the "end of history," in which Hegel's dialectic culminates. Theory must dispense with the search for a stable starting point from which investigation of the artwork proceeds. This has become the case following Nietzsche's dismantlement of the truth of metaphysics in his revelation that "the 'apparent' world is the only true one: the 'true' world is merely added by a lie."[8] Heidegger's reflections on Nietzsche illustrate the world philosophy must now address:

> [I] If the world were constantly changing and perishing, if it had its essence in the

[4] Kant 18.

[5] *Ibid*. 225.

[6] *Aesthetic* 18.

[7] *Ibid*. 181.

[8] Nietzsche 481.

> most perishable of what perishes and is in-constant, truth in the sense of what is constant and stable would be a mere fixation and coagulation of what in itself is becoming: measured against what is becoming, such fixation would be in-appropriate and merely a distortion… A knowledge that – as true – takes something to be "being" in the sense of constant and stable restricts itself to beings, and yet does not get at the actual: the world as a becoming world.[9]

A world of becoming is one in which philosophy cannot content itself with the stability of the apparent. The notion of a stable reality becomes mythical. Rather, reality is disunified, fragmented, constituted by the sedimentation of power and history. Historical contingency resists the thrust toward universality which motivated the aesthetics of Adorno's predecessors. Adorno's project reorients this thrust in philosophy, as he writes, "The great philosophical aesthetics stood in concordance with art to the extent that they conceptualized what was universal in it; this was in accordance with a stage in which philosophy and other forms of spirit, such as art, had not yet been torn apart."[10] Aesthetics must now attune itself to the processual and therefore fragmented nature of reality, within which the subject engages the work of art, and through which philosophy must position itself.

Hence, "art" is neither a stable category nor a catalog of exemplary works. Aesthetics cannot begin with reflections on art, but must ground itself in the individual artwork. Indeed, art does not exist apart of a world of becoming, which only knows individual works. Moreover, the work itself is not anything stable, whose value and meaning transcend historical interpretation. "The artwork is a process essentially in the relation of its whole and parts. Without being reducible to one side or the other, it is the relation itself that is a process of becoming." Whatever the theorist labels the totality of the work cannot be a "structure that integrates the sum of its parts."[11]

The immediate consequence for aesthetics is this crisis of art's self-identity. If the post-metaphysical worlds strips artworks of their ideas, aesthetics cannot aim to reach behind the work to capture its truth. Aesthetics must critically engage, and not blindly surrender itself to a fractured reality. The relationship between subject and work is not immediate, nor can philosophy penetrate what truths hold in this relation. Aesthetics must mediate the intersection of work, society, and history. This mediation is necessary because no aesthetics grounded in a systematic conceptual apparatus can do justice to the individual work. If the world is becoming, then theories of aesthetics must relinquish their reliance upon systematicity; to understand the work is not the same as pumping it full of philosophic concepts. In short, form must give way to experience. Only the latter is equipped to address art in a world of becoming:

> The exertion of cognition is predominantly the de-struction of its usual exertion, of its using vio-lence against the object. Knowledge of the object is brought closer by the act of the subject rending the veil it weaves about the object. It can do this only when, passive, without an-xiety, it entrusts itself to its own experience. In the places where subjective reason senses subjective contingency, the primacy of the object sh-immers through; that in the object which is not a subjective addition.[12]

While the "subject is the agent," aesthetics cannot allow it to be "the constituent of object."[13] The requirement that the subject rend "the veil it weaves about the object"

[9] Heidegger 64.

[10] *Aesthetic* 334.

[11] *Ibid.* 178.

[12] *Critical* 254.

[13] *Ibid.*

demands more than that the abdication of aesthetic systems. The subject, too, finds itself interwoven within the forces of power and history. Thus a dialectical aesthetics performs the requisite task by submitting the reification of systematic thought to critical revision. It must resist positivity in its construction of concepts, and dispense with identity thinking which represents the object in the image of the concept. The violence done by the concept compels Adorno to posit as a criterion of an aesthetics' success the capacity to draw from the artwork a critical revision of our representation of reality.

This imperative to advance socio-historical critique heightens against the backdrop of the universal instrumentalism late capitalism imposes upon life. Aesthetics must not only pry itself loose from the rigidity of conceptual systems, but also save the work of art from the spell of its commodification. I would like to suggest that this is the central objective of *Aesthetic Theory*, and the work of the Frankfurt School generally: resistance to the valueless fungibility we face in a world circumscribed by instrumental rationality. The latter takes as its operating principle the reduction of all aspects of life, including art, to their use value. Under these conditions, the work of art is reduced to a unit of pure exchangeability, a commodity circulated in the market. As a result, Adorno recognizes that "art no longer has a place" in our society. Under instrumental reason, "art fragments on one hand into a reified, hardened cultural possession and on the other into a source of pleasure that the consumer pockets and that for the most part has little to do with the object itself."[14]

The modern era marks the culmination of rational-Enlightenment thought, whereby the empirical world succumbs to the Kantian transcendental subject – object becomes subject. As a consequence, "thought makes itself mere tautology."[15] Late capitalism embodies the socio-economic concretization of instrumental reason, which reifies consciousness as identity thinking. The market, through its rational-economic modes of thought, has seized the subject from the world, thereby neutralizing the subject's critical relationship to the world. The subject now unknowingly becomes the object of a world in which one-dimensional thought dominates. This makes the subject increasingly unable to perform the task Adorno demands of aesthetics, to subject reality to critical revision through the artwork, to see the world otherwise. Instead, everywhere the subject goes, it confronts only itself: "The man of science knows things to the extent that he can make them. Their "in-itself" becomes "for-him." In their transformation the essence of things is revealed as always the same, a substrate of domination."[16] The complete administration of society instrumentalizes all spheres of life to such a degree that knowledge hypostasizes into the operations of a machine. Experience has completely given itself away to form. Adorno takes these consequences as the central target to which dialectical thinking must respond, which lays the foundations of negatively dialectical thinking:

> To grasp existing things as such…to think of them as surface, as mediated con-ceptual moments which are only fulfilled by revealing their social, historical, and human meaning – this whole aspiration of knowledge is abandoned. Knowledge does not consist in mere perception,class-ification, and calculation but precisely in the determining negation of whatever is directly at hand. Instead of such negation, mathematical formalism, whose medium,

[14] *Aesthetic* 15.

[15] *Dialectic* 20.

[16] *Ibid.*

> number, is the most abstract form of the immediate, arrests thought at mere immediacy.[17]

Adorno, in collaboration with Max Horkheimer, thus crystallizes the project of a dialectic aesthetic theory. This notion of dialectic, however, differs from its original Hegelian and Marxist variations. Dialectics must be negative. This means it neither submits itself to the positivity of idealism's synthesis (the sublation of thesis and antithesis), nor does it operate according to the objective laws of historical materialism. Instead, dialectic thought must dwell in the "determining negation of whatever is directly at hand." It must take as its object the concretization of a fragmented and antagonistic reality.

It is from the objectives set forth by Adorno's critical project that I take this essay's investigative point of departure. *Aesthetic Theory* offers a response to a completely administered post-metaphysical world – it presents a dialectical aesthetics that dislodges the artwork from its social appropriation by instrumental rationality. The question I am posing is how does the philosophical work, in particular that of Adorno, respond? How does the philosophical text resist its appropriation by a reified world that strips objects of their inherent value, reduced to market commodities. My own experiences as a young philosophy student are illustrative of the predominance instrumental reason has claimed over the modern era. Indeed, it is difficult for me to recall an instance where I revealed my major and did not receive the surefire response, "what are you going to *do* with that?"

Admittedly, unlike art, philosophy is not so much an exchanged commodity. The humbling truth is that it is more commonly confined within the walls of the academy. But nonetheless, I am interested in asking how the philosophical work resists the spell of instrumentality that has seized our world. I will argue that *Aesthetic Theory* provides an answer in its philosophic style, achieved though a constellational and negatively dialectical construction. Adorno does this by modeling his philosophy in aesthetic experience. This is not to say that *Aesthetic Theory* is an artwork, but that it integrates elements unique to art that preserve its critical capacity. These elements, however, cannot be brought out independent of the text's internal development of its ideas. *Aesthetic Theory* is not a work of literature that subjects itself to aesthetic analysis. The text offers a theory that unfolds through its manner of presentation, but the theory reflexively shapes the text's presentation. Thus, my investigation will begin by drawing from the text what it means for a philosophical work to resist its instrumentalization. I will first argue that the work must overcome its utility as a method, from which I will uncover how *Aesthetic Theory* performs this task.

II. Artworks retain a critical dimension given their situation both within and outside the world. This is what distinguishes artworks from inert objects - their resistance to the world within an explanative context. Adorno writes, "Only by virtue of separation from empirical reality…does the artwork achieve a heightened order of existence."[18] Great artists Adorno looks to, such as Rembrandt, Beckett, or Beethoven, were among those whose "sharpest sense of reality was joined with estrangement from reality."[19] Hence the necessity for a dialectical aesthetics arises from what is artful in the work. For aesthetics to tend to the work, it must preserve art's autonomy negatively – its own negative participation within reality:

[17] *Ibid.*

[18] *Aesthetic* 4.

[19] *Ibid.* 9.

> "That art on the one hand confronts society autonomously, and, on the other hand, is itself social, defines the law of its experience."[20]

But how can aesthetics engage a work that is both within and outside reality, given the position of the theorist within society? The subject must understand the work as an antagonistic movement between its inner parts. If the work's own determination opposes reality, it cannot be understood as a self-enclosed identity. A world of becoming precludes the possibility of aesthetic comprehension that exhausts the work's meaning. Rather, "a relation, not identity, operates between the negativity of the metaphysical content and the eclipsing of the aesthetic content."[21] The question thus persists, a relation between what? The traditional aesthetic categories such as form-content or universal-particular cannot capture this relation; they must give way to an experiential aesthetics. No framework of binaries can capture the antagonistic *processes* in the work, as Adorno specifies:

> Whatever may in the artwork be called totality is not a structure that integrates the sum of its parts. Even objectified the work remains a developing process by vir-tue of the propensities active in it. Conversely, the parts are not something given, as which analysis almost inevitably mistakes them: Rather, they are centers of energy that strain toward the whole on the basis of a necessity that they equally perform. The vortex of this dialectic ultimately consumes the concept of meaning. [22]

That the parts of the work are "not something given" forecloses their capacity to ground an aesthetics. There exists nothing for the subject to grasp a hold of in the work, despite instrumental reason's claim to do so. Adorno instead couches the work's elements in a discourse of becoming: "developing process," "centers of energy," "vortex of dialectic." But what allows Adorno to assert such claims about the artwork? Could we not say that he has injected the work with a dialectic method of process, thereby sacrificing experience to form?

Quite the contrary, Adorno's account presupposes an aesthetic experience on the part of the reader without proactively pointing to a particular work. The philosophy we read does not take its object as given. It suspends reliance upon ground and thereby engages a world of becoming. All that Adorno can rely upon is the experience of artworks, and not solely what is determinate within them. Hence he turns to our experience, upon which he invites us to reflect. What Adorno points out is what he sees taking place as he experiences the work, as if to say, "there it is; do you see it too?" In short, *Aesthetic Theory* offers a philosophy of reflection and not conceptual projection. This is what allows Adorno to posit the sorts of reflective identity statements we continuously find in the text: "The artwork is X, art does Y," and so forth.

This means that Adorno dispenses with the task of demonstration. He does not begin with the work's elements and move outward in order to identify the work any more than he injects the work with concepts external to

[20] *Ibid.* 348.
[21] *Ibid.* 358.
[22] *Ibid.* 178.

it. The movement is not unidirectional, but mediates what is experienced in the work. This is confirmed by the position of the above passage within the text. It rests between a preceding paragraph that spans three pages, a fragmented discussion of Mozart, Beethoven, "Stockhausen's concept of electronic works," and "Picasso's rayonism," and an ensuing account of aesthetics' relation to Kant and Stravinsky. The passage is a brief moment of clarity caught within a discussion whose movement is suggestive of the artwork's antagonistic elements. Adorno's insight is not the conclusion of a logically developed argument but an instant of reflection that allows the artwork's internal friction to shine.

Adorno's style allows his dialectic aesthetics to reflect on an experience of the artwork. The text's construction is not a hierarchical presentation of concepts; it instead subjects itself to dialectic thought. Thus *Aesthetic Theory* models itself in aesthetic experience. Indeed, if the artwork is a movement of antagonisms, then aesthetics' response must attune itself to this processual experience. Aesthetics cannot rigidify itself anymore than the work of art can abdicate its internal movement. Thus art calls for an aesthetics that is dialectical, and thereby allows the subject to engage the work without appropriating it:

> To whoever remains strictly internal, art will not open its eyes, and whoever remains strictly external distorts artworks by a lack of affinity. Yet aesthetics becomes more than a rhap-sodic back and forth between the two standpoints by dev-eloping their reciprocal mediation in the artwork itself.[23]

This means the subject's consciousness must "remain constantly mobile both internally and externally to the work."[24] We witness this mobility unfold through the stylistic strategies Adorno deploys in the development of his theory. The text's ideas linger, they do not explain away the meaning of either art or their own theory. What makes the latter in fact dialectical is not just its capacity to reflect upon the particularity of the work in its non-identity; a dialectical aesthetics must maintain its coherence through its own dialectic mediation. And this occurs in the fragmentary style in which *Aesthetic Theory* grounds its aesthetics.

A dialectical aesthetics thus jettisons conceptual systems in order to afford the work its autonomy. This allows the work to exist apart from the world that struggles to pull it back. Instrumental reason endeavors to harden the work in order to fetishize its value as a commodity within the market. At the same time, a post-metaphysical world of becoming renders conceptual aesthetic methods untenable. Adorno recognizes, "the tendency of philosophical aesthetics toward those abstract rules in which nothing is invariable... is transient; the claim to imperishability has become obsolete."[25] Yet what prevents the petrifaction of Adorno's dialectic aesthetics into "abstract rules"? In other words, how does an aesthetics, amid an instrumentalized world, ward off its conversion into method, thereby offering itself as a dispensable tool to the art critic?

This concern is eminent for Adorno, who warns, "the over-valuation of method is truly a symptom of the consciousness of our time...this tendency is related to the nature of the commodity: to the fact that everything is seen as functional, as a being-for-another" (Goldmann 129). Thus, dialectical aesthetics must turn its negation back on itself. That is, it must make a concerted effort to resist reliance upon conceptual schemata in order

[23] *Ibid.* 350.
[24] *Ibid.*
[25] *Ibid.* 339.

to understand the artwork. Yet Richard Wolin suggests *Aesthetic Theory* fails in this task. Wolin asserts that Adorno's aesthetics "remained undialectically wedded to the concept of an esotericized, autonomous art as an absolute model of aesthetic value."[26] Consequently, *Aesthetic Theory* "runs the risk of a false sublation of autonomous art, whereby a crucial refuge of negativity and critique would be prematurely integrated with facticity as such."[27] Yet Wolin's claims that Adorno remains "undialectically wedded" to autonomous art undervalues the critical comportment of such art, and moreover, Adorno's incorporation of similar elements into his own work. I would like to suggest that criticism such as Wolin's neglects this latter point, the critical style of *Aesthetic Theory*, which dialectically preserves its coherence in the face of its instrumentalization. It is toward this insight that I direct the following section.

III. Nowhere in *Aesthetic Theory* do we find a definition of what it means for an aesthetics to be negatively dialectical, at least not one that exhausts the multiplicity of its dimensions. Nor do we find examples that demonstrate the theory's proper application. This feature secures the complexity of the work, its multiple layers and hypnotic affect in its series of digressions and philosophic excursions. But rather than dispel the force from the text, these qualities ensure its successful resistance to its instrumentalization as a method.

Aesthetic Theory, in short, is enigmatic, a term Adorno uses to describe the artwork's autonomous position both within and outside society. But the artwork's autonomy, its internal antagonistic movement, does not reduce it to a unit of chaos. Rather, within its enigmaticalness the subject encounters the artwork's critical dimension. This is what allows Adorno to contend, "the idea of a conservative artwork is inherently absurd." This is because the artwork occupies a critical posture beyond the limits of its social inception,

> "By emphatically separating themselves from the empirical world, their other, they bear witness that that world itself should be other than it is; they are the unconscious schemata of that world's transformation."[28]

How is it that the work's enigmaticalness, in its dynamic presence in and absence from this world, critically negates society? Adorno's answer is its truth-content, which the work possesses as its own cognitive capacity, what allows the work to remain an object not subsumable by the subject's concepts. The work's truth-content orients the movement of its internal parts, which depends upon philosophy for its self-actualization. Works aim toward the "determination of the indeterminate" in their resistance to reality, but in so doing they simultaneously pose a problem, that of their negative dimension. This is why we do not look at artworks and immediately think "Revolution!" Instead, the work "achieves meaning by forming its emphatic absence of meaning."[29] No interpretation will reveal the work "as a new immediacy" because the work's "enigmatical-ness outlives the interpretation that arrives at the answer."[30] Thus it becomes the objective of a dialectical aesthetics to no longer "explain away the element of incomprehensibility" but instead "understand the incomp-rehensibility itself."[31]

[26] Wolin 45.
[27] *Ibid.*
[28] *Aesthetic* 177
[29] *Ibid.* 127.
[30] *Ibid.* 125.
[31] *Ibid.* 347.

In order to perform this task, the subject allows itself to be disciplined by the truth-content of the work. Thus it becomes the task of aesthetics to reflect on the work's truth: "By demanding its solution, the enigma points to its truth-content. It can only be achieved by philosophical reflection. This alone is the justification of aesthetics."[32] Philosophy becomes the midwife of the work's truth; it allows the unreality of the work to demonstrate the inadequacy of what is real. But this only occurs when the subject allows the work to speak for itself, in other words, how to be fully autonomous. Philosophy must dispense with its metaphysical quest for truth and submit to the truth of the work. Martin Jay clarifies, "Truth for Adorno was not... merely correspondence between propositions and an external referent in the current world, but rather a concept with normative resonances as well, referring to a future 'true' society."[33] The question thus arises, how does a dialectical philosophy perform this task insofar as there exists nothing stable within the work to comprehend?

Philosophy draws forth from the work its suspension of what is given in the world. Philosophy takes up a reflective task that is experiential; it provides, through the work, a moment of insight into reality's indeterminate other. It is the artwork which provides an antipode of the concept toward which dialectic thought moves. As the subject lingers in the work's truth-content, it makes manifest the violence of the philosophical concept. A dialectical aesthetic experience paralyzes the subject-object distinction, and in its paralysis the concept confronts its own inadequacy. This inadequacy is what submits reality to the possibility of its other; philosophical reflection on the work's truth-content interweaves reified consciousness and reflective self-consciousness, thereby estranging the subject from its hypostatized relation to reality.

This experience re-orients the way in which the subject engages the object. Philosophy does not content itself with the grasping of concepts, but instead realizes itself as an experience. Hence, *Aesthetic Theory* presents itself as a reflective exercise and not a mere handbook of aesthetics. A dialectical aesthetics offers a "how" of thinking and not a "what," which is to say philosophy invites the subject to follow a *way* of thinking. Philosophy becomes a verb instead of a noun. It is under this light that we should understand *Aesthetic Theory* as a text that moves its reader. Indeed, the text lingers in the experience of its object, the artwork, and therefore becomes artful in its movement, much like a musical score. In so doing, the text resists its instrumentalization; its thought depends upon our active participation. The way of thinking upon which it drafts the reader is not a method. It is slippery, antagonistic, much like the engimaticalness of the artwork, yet remains distinct from art in its presentation of concepts. We can trace this movement in the following passages:

> Even by artworks the concrete is scarcely to be named other than negatively. It is only through the nonfungibility of its own existence and not through any special content that the artwork suspends empirical reality as an abstract and universal functional nexus. Each artwork is utopia insofar as through its form it anticipates what would finally be itself, and this converges with the demand for the abrogation of the spell of self-identity cast by the subject. No artwork cedes to another.[34]

Only against the suffering of reality does the work revive its own singularity, its nonfungibility, and thereby direct its movement toward utopia. In *Aesthetic Theory*, philosophy responds by negatively presenting

[32] *Ibid.* 128.

[33] Jay 159.

[34] *Aesthetic* 135

the contradictory movement of the work. The text does not explain away this movement, but participates in it. Indeed, the "artwork is utopia," yet simultaneously utopia is not entirely of the work; it is anticipated and dependent upon the subject's "abrogation of the spell of self-identity." We follow philosophy's mediation of the artwork in concurrence with its own account of the work. The demand the artwork places on philosophy is simultaneously fulfilled by the dialectical style of the text. This movement continues through the passage:

> The nonfungibility, of course, takes over the function of strengthening the belief that mediation is not universal. But the artwork must absorb even its most fatal enemy – fungibility; rather than fleeing into concretion, the artwork must present through its own concretion the total nexus of abstraction and thereby resist it.[35]

The text shares a movement of resistance with the artwork it portrays. The artwork's absorption of fungibility folds within the "nonfungibility" of its "own existence." Deep within the artwork's "nexus of abstraction" and resistance to it, dwells the call for a mode of thinking that pleasures before non-identity. *Aesthetic Theory* heeds this call as it unfolds through a "sequence of dialectical reversals and inversions."[36] Much like the aesthetic experience, Adorno's philosophy is present yet retreats from itself; it develops in a manner that is non-identical. Indeed, the movement of the text mediates between its aesthetics and glimpses into the individual works, which we can see in the continuation of the passage:

> Repetition in authentic new artworks is not always an accommodation to the archaic compulsion toward repetition. Many artworks indite this compulsion and thereby take the part of...the unrepeatable; Beckett's *Play*, with the spurious infinity of its reprise, presents the most accomplished example. The black and grey of recent art, its asceticism against color, is the negative apotheosis of color.[37]

The works mentioned do not serve to demonstrate a theory, nor does the brevity of their description undermine their relevance to Adorno's aesthetics. They belong to the movement of the text, which rides their aesthetic experience. This movement is confirmed in the following moment:

> But because for art, utopia – the yet to exist – is draped in black, it remains in all its mediations recollection; recollection of the possible in opposition to the actual that suppresses it; it is the imaginary reparation of the catastrophe of world history; it is freedom, which under the spell of necessity did not – and may not ever – come to pass. Art's methexis in the tenebrous, its negativity, is implicit in its tense relation to permanent catastrophe.[38]

The textual movement of *Aesthetic Theory* slides from its aesthetics to particular works and back to aesthetics in a seemingly fragmentary manner. Much like the artwork, the text feels tenebrous in its negativity, yet resists its own catastrophe by its very negativity. It does so in its negatively dialectical movement, which grounds itself in the experiences of artworks, such as Beckett's *Play* or "black and grey" ascetic art. The experience of particular works offers itself as the only foundation for an aesthetics due to the conditions philosophy faces in a post-metaphysical world that dismantles all stable foundations. Hence, the text moves in a non-deductive manner; it submits its own grounds to dialectical reflection. This means the text resists its own given-ness. Though constricted to the confines of the pages on which it is written, the text reads temporally. It

[35] *Ibid.* 135.
[36] Richter 101.
[37] *Aesthetic* 135.
[38] *Ibid.*

moves like a musical composition, but not like a pop song or Jazz piece, both of which Adorno disdained for their rationalized predictability. The text's form does not dictate the ordering of its contents, but instead, like an atonal Schoenberg composition, it weaves through the friction of its own elements.

This is not to say that philosophy implodes in its resistance to the demands of a world circumscribed by instrumental reason. The text does not forego rigorous philosophical investigation, but delicately mobilizes the concept through its own self-effacement. This strategy is made possible by the text's constellational style. The constellation ties together conceptual moments in such a way that resists the Kantian distinction between subject and object. The style of *Aesthetic Theory* bridges the chiasmic gap, as Kant figured it, between phenomena and noumena. By grounding itself in aesthetic experience, the text unfolds through constellations that give coherence to a dialectical aesthetics composed of conceptual formations that retain empirical phenomena. Concepts do not subsume the particular; the latter sustains itself through the former. Susan Buck-Morss clarifies this epistemological strategy:

> Cognitive knowledge... was achieved by means of abstraction: the particular entered into the concept and disappeared. But in [constellations] the particulars, although conceptually mediated, reemerged in the idea...they *became* the idea in the conceptual arrangement of their elements. The role of the subject, to draw connections between the phenomenal elements, was not unlike that of the astrologer, who perceived figures in the heavens.[39]

In a post-metaphysical world, in which there are neither things-in-themselves nor transhistorical ideals, the subject's encounter with the world is limited to its particular empirical phenomena. The constellations that comprise *Aesthetic Theory* retain aesthetic experiences as non-hierarchical monads in an inter-connected web. Each moment, like a single star of a constellation, contains the totality, its own picture of the world, yet remains distinct from the other moments.

The constellated form of *Aesthetic Theory* allows the text to adapt itself to a world of becoming in that its constellations do more than simply present what is empirically given in artworks. The empirical task is the role of science, which submits empirical facts to research. Rather, each moment of the text interprets the fragmentary reality constitutive of the given. These interpretations manifest mediation at work; they penetrate the historical contingency and socially constructed dimensions of reality. Recall that mediation, for Adorno, is negative and in opposition to the Kantian conceptual appropriation of the object. Adorno's mediation of artworks in *Aesthetic Theory* preserves non-identity, and in so doing unwinds particulars from their conceptual reification. Unlike the aesthetics of Kant and Hegel, Adorno's does not ground itself in unity, as confirmed by the fragmentary and non-narrative structure of the above passages. This explains how the text does not offer an aesthetic method, but instead submits itself to the particularity of the artwork. What we find is an aesthetics attuned to the fragmented and contradictory nature of reality in bourgeois society. Where instrumental reason congeals this realm of contradictions into systems of identities, *Aesthetic Theory* unfolds through constellations that render visible what is antagonistic of reality. In a world that converts critical thinking into a method, dialectical aesthetics responds in its resistance to commodification. In the

[39] Buck-Morss 92.

market, commodities operate according to the principles of abstraction, identity, and reification, which Buck-Morss describes as the "ossification of the object as a mystifying fetish by splitting it off from the process of its production."[40] In contrast, dialectical aesthetics reunites what instrumentality splits.

IV. The truth of the artwork lies beyond reality in its estrangement from the society in which we receive it. But how does the work retain its negativity, its openness to a world that does not yet exist despite its existence within a reified world? It is in mimesis that art's truth-content remains in contact with a world other than the empirical: "By pursuing its own identity with itself, art assimilates itself with the nonidentical: This is the contemporary stage of development of art's mimetic essence."[41] The work's "mimetic essence" is the object of philosophy's reflective response to the work. It makes possible the work's openness to a primal world before the subject's abstraction from nature through reason. Mimesis is what cannot be spoken in the work; it is a mystic openness to a world not translatable into language.

What cannot speak in the artwork, and therefore invites philosophical assistance, is its mimetic comportment, which Robert Kaufman describes as that which is "grasped not as transcription but as an attempt provisionally to know something of the otherness outside the subject."[42] This otherness resists conceptual appropriation because it belongs to a world from which the concept has not yet abstracted itself. Frederic Jameson clarifies that mimesis "can be said often to function as a more adequate substitute for the primal relationship of subject and object."[43] What is primal of mimesis "forestalls dualistic thinking by naming the dualism as such."[44] Thus, mimesis works against the concept, forcing it back on itself. Rather than appropriate the work, the concept must approach the work by way of dialectical reflection – it must experience that which cannot be named using the subject's tools of reason.

The nature of mimesis is one of non-identity, which occupies a world that is both primal and of the future. On the one hand, mimesis belongs to a world that exists prior to the ascendance of the bourgeois subject and its domination of nature. Adorno writes, "Art is imitation exclusively as the imitation of an objective expression, remote from psychology, of which the sensorium was perhaps once conscious in the world and which now subsists only in artworks."[45] In its mimetic dimension, the artwork offers a glimpse into what subjective consciousness has alienated from nature. Yet on the other hand, mimesis opens a free world not yet realized, beyond the subjective logic of identity:

> Only the autonomous self is able to turn critically against itself and break through its illusory imprisonment. This is not conceivable as long as the mimetic element is repressed by a rigid aesthetic superego rather than the mimetic element disappears into and is maintained in the objectivation of the tension between itself and its antithesis.[46]

The artwork's mimetic comportment, which Adorno goes on to describe as "the plenipotentiary of an undamaged life in the midst of mutilated life," is not recuperated in a nostalgic past, but imitates a world not yet actualized within our own. Thus, mimesis

[40] *Ibid.* 98.
[41] *Aesthetic* 134.
[42] Kaufman 201.
[43] Jameson 105.
[44] *Ibid.* 105.
[45] *Aesthetic* 112.
[46] *Ibid.* 117.

assumes different contexts, both of the past and future, throughout the text. *Aesthetic Theory* does not advance an extractable notion of mimesis; its account, instead, can be said to unfold mimetically. That is to say mimesis works against the constraints of language. Our understanding of mimesis develops in fragments, each of which unfold within the contours of a particular constellated moment.

The text's fragmented account of mimesis illustrates its constellational structure and movement. In certain moments, mimesis is that which is remembered in the artwork; at other moments it speaks to the emancipatory potential of the work. These accounts are, of course, not mutually exclusive, as Adorno states, "The trace of memory is mimesis, which every artwork seeks, is simultaneously always the anticipation of a condition beyond the diremption of the individual and the collective."[47] In it's fragmented and divergent moments, mimesis unfolds through the constellations in which its various account are situated. The text charges mimesis with different social, historical, and aesthetic valences, but never does so in a univocal manner.

> I would like to suggest that *Aesthetic Theory* contains a mimetic dimension that allows it to resist its instrumentalization. Mimesis allows the object to speak for itself and therefore counters the violence done to it by subjective consciousness. It occupies a purely experiential world that escapes its petrification in language. Yet we cannot say the text is mimetic insofar as its medium is language, which necessarily obfuscates what is intrinsically unspeakable in mimesis: "By virtue of its double character, language is a constituent of art and its mortal enemy." Adorno goes on to clarify that "compared to significative language" the expression of mimesis "is older though unfulfilled."

Mimesis, what affords the artwork its dimension of negativity, resists its inclusion within philosophy. Yet philosophy depends upon a mimetic dimension in order to preserve the primacy of the object apart from its conceptualization. Philosophy thus confronts a paradox: in order to resist its instrumental concretization it must not subsume its object, yet this requires mimesis, which resists linguistic translation. Thus philosophy must work against language by way of language, and this is what *Aesthetic Theory* accomplishes through its constellational style.

The text demonstrates its resistance to language in its constellational account of mimesis. As we have seen, mimesis unfolds in fragmentary moments. Each moment belongs to a constellation, yet no moment provides the complete picture. These moments are self-contained, yet simultaneously bleed into one another. Contradictory accounts are assembled alongside each other. They are not smoothed out and narrated, their development successive, but rather manifest the work of conceptualization. As disunited assemblages, the divergent accounts of mimesis expose a glimpse of a world in which thought lingers yet does not conceptualize its object.

There exists within the text an account of mimesis, its idea moves through the text's constitutive constellations, yet it never concretely presents itself before us. In other words, *Aesthetic Theory* employs language to provide an account of mimesis, but never names it directly. Language is all that we read, but language never identifies mimesis. Instead, constellations invite us to perform this task as we piece together their fragmentary moments. Consequently, we proceed through the text in a dialectical manner. It's idea of mimesis is absently present; it is there, but not in its self-identity.

[47] *Ibid*. 131.

Philosophy assumes a mimetic comportment of its own in its attempt to offer an unknowable alternative to the world it critically negates. Thus philosophy mobilizes itself against itself: it becomes dialectical not only in its content, but also in the style through which its content develops. *Aesthetic Theory* is certainly philosophic. But it preserves itself, in resistance to its instrumentalization as method, in its opposition to the concept, upon which philosophy has traditionally relied. In the face of a post-metaphysical world, the text presents itself as an experience, which has otherwise been denied by the grip in which instrumental reason binds our world.

Works Cited

Adorno, Theodor. *Aesthetic Theory.* Trans. Robert Hullot-Kentor. Minneapolis: University of Minnesota Press, 1997.

Adorno, Theodor. *Critical Models: Interventions and Catchwords*. Trans. Henry W. Pickford. New York: Columbia University Press, 2005.

Buck-Morss, Susan. *The Origin of Negative Dialectics: Theodor Adorno, Walter Benjamin, and the Frankfurt Institute*. New York: The Free Press, 1977.

Goldmann, Lucien. *Cultural Creation*. Trans. Bart Grahl. Oxford: Basil Blackwell & Mott Ltd, 1977.

Heidegger, Martin. *Nietsche vol. 3 &4*. Trans. Joan Stambaugh, David Farrell Krell, Fran A. Capuzzi. Ed. David Farrell Krell. New York: Harper & Row, 1982.

Horkheimer, Max and Theodor Adorno. *Dialectic of Enlightenment*. Ed. Gunzelin Schmid Noerr. Trans. Edmund Jephcott. Stanford, California: Stanford University Press, 2002.

Jameson, Frederic. *Late Marxism: Adorno, Or, The Persistence of the Dialectic.* New York: Verso, 1990.

Jay, Martin. *Adorno*. Cambridge: Harvard University Press, 1984.

Kant, Immanuel. *Critique of Judgment*. Trans.

Kaufman, Robert. "Lyric's Expression: Musicality, Conceptuality, Critical Agency." *Cultural Critique* 60 (2005): 197-216.

Nietzsche, Friedrich. "Twighlight of the Idols." Trans. Walter Kaufman. *The Portable Nietzsche*. Ed. Walter Kaufman. New York: Penguin Books, 1982. 463-564.

Richter, Gerhard. "Aesthetic Theory and Nonpropositional Truth Content in Adorno ." New German Critique 33 (2006): 119-135.

Wolin, Richard. "Utopia, Mimesis, and Reconcilliation: A Redemptive Critique of Adorno's Aesthetic Theory." *Representations* 32 (1990): 33-49.