

TABULA RASA

RUTGERS UNIVERSITY
UNDERGRADUATE PHILOSOPHY JOURNAL

ISSUE 1, SPRING 2008

ACKNOWLEDGMENTS

The Staff of the Rutgers Philosophy Journal would like to thank the Rutgers Philosophy Department, especially Mercedes Diaz, for making this Journal a reality. We would also like to thank all the authors who submitted papers.

All contact information can be found on the Rutgers Philosophy Journal Website.

Special thanks go out to Daniel Schaub for contributing his expertise in web development in making our website really pretty.

The website can be found at the following URL:
<http://philosophy.rutgers.edu/UNDER/JOURNAL>

TABULA RASA

EDITOR-IN-CHIEF

Paul Chiariello

EDITOR-OF-CONTENT

Max Bialek

PUBLISHER

Kamil Kaczynski

SECRETARY

Rachel Saitzyk

FACULTY ADVISOR

Martin Lin

EDITORIAL STAFF

Sophie Ban	David Infortunio
Eugene Belenitsky	Andrew Kobylarz
Greg Bueno	Max Mintz
Uriel Carni	David Morgan
Gilda Charles	Cory Nichols
Helen Ciacciarelli	Zachary Perry
Nicholas Farmer	Nina Ren
Summer Faruqi	Alina Rybakov
Chase Freed	Barbara Saramak
Steve Gallagher	Nicholas Slininger
E.J. Green	Volodymyr Takhistov
Hillel Herzfeld	Henry Yeh
Andrew Hurwitz	

LETTER FROM THE EDITORS

The responsibilities of a philosopher in academia nowadays are many and varied: teaching, entertaining guests, advising students, departmental duties, and sometimes even philosophizing. An experienced professor has done all these things, and can probably name what they love most and least about their jobs. Before a professorship, one spends their time as a graduate student. The scope of their responsibilities is similarly large, and any graduate student (proud or bitter!) would be more than happy to tell you what they do and do not like about the path that they have set for themselves. But, before any of us reach those stages in life, we are wide eyed undergraduates. Every glimpse into the world that we hope to one day join, from the classic papers we may read to the mundane routines of our professors, excites and inspires us. To edit and review the work of their peers as part of a journal is one of those responsibilities that our professors take on, and it was to get a first taste of what we hope life might be like in the future that we decided to create just such an opportunity for ourselves.

We received over thirty submissions from an international collection of colleges, and we are happy to present you with what our editorial staff picked out as the best five. From our home of Rutgers—New Brunswick, Caleb White offers an interesting exploration of the way “mind enhancing drugs” might effect how praiseworthy or blameworthy we would be as moral agents. Two fantastic papers were received from across the Atlantic. Konstnacja Duff, from Cambridge, looks at the political philosophy of Hobbes, and asks at what point our obligations to the body politic end and the option of resistance or rebellion becomes justifiable. Also from Cambridge, William Crouch draws on Wittgenstein’s work *On Certainty* in an effort to resolve the problems of skepticism. Our final two papers are from students at Princeton, just one train stop away from Rutgers. Wesley Bronson examines the role that the intentions of an agent play as part of the necessary and sufficient conditions for moral responsibility. Last, but not least, the brain in a vat hypothesis is analyzed by Avi Miller, and a distinction between truth in the natural and truth in the artificial is developed.

We hope you enjoy the papers to be found in this inaugural issue of the Rutgers Undergraduate Philosophy Journal. We greatly enjoyed producing it, and look forward to working on its future publications.

TABLE OF CONTENTS

Responsibility Under Drug Induced Mental Enhancement

Caleb White

1

Hobbes and the Limits of Political Obligation

Konstancja Duff

21

Wittgenstein on Scepticism in ‘On Certainty’

William Crouch

33

Intentionality, Intending and Moral Responsibility

Wesley H. Bronson

43

The Natural-Artificial Distinction:

A Limit Concept for Truth and Reality and its Application Across Matrices

Avi M. Miller

59

Responsibility Under Drug Induced Mental Enhancement

Caleb White
Rutgers University

This article examines the issue of whether and to what degree a person is praiseworthy for any accomplishments she achieves while under drug-induced mental enhancement, not the permissibility of using any such drugs. I argue that the use of mind-enhancing drugs does not limit or reduce the praiseworthiness of an agent to any degree. I begin by differentiating three types of mental enhancements (mental stamina, memory, and ‘cognitive capacity’). Then I point out that mental enhancement is different from mental creation, and conclude that agents with drug-induced mental enhancement, but not creation, are fully praiseworthy. After this we reach my main thesis, which is that the basis for intellectual praise is in one’s preexisting mental states and mental labor. Since enhancement drugs cannot affect one’s preexisting mental states and does not affect the agent’s connection to her mental labor, I conclude that these drugs do not affect praiseworthiness. Finally I address certain possible objections to this conclusion and present my own counterarguments.

SECTION I

Introduction

Imagine, if you will, your favorite philosophical or literary work. Now imagine that its author produced this masterpiece while using some sort of mind- (or intelligence-) enhancing drug. Do you like it less? Are its theories or stories less brilliant? Perhaps most importantly, does the author deserve less praise? Or any praise at all? The reality is that many new mind-enhancing drugs are being discovered or created, not to mention the already increasing use of so called “study drugs” (often amphetamines) which are currently available. Students of all ages are being prescribed or illegally using drugs that are specifically designed to increase one’s ability to perform mental tasks, but how this affects the praiseworthiness of these people has been largely overlooked. Thus it is of practical importance, in this new and rapidly advancing atmosphere of enhancement drug use, to examine the situation and decide how to treat and regard the people who use them.

This paper will discuss an agent’s *responsibility* (praiseworthiness) for accomplishments she achieves while under the influence of mind-enhancing drugs, *not the permissibility* of using any such mind-enhancing drugs. *Enhancement* will be a key term throughout this paper and should be understood strictly to involve the alteration of the preexisting condition of the agent. Any drug which creates totally new and distinct qualities or characteristics of an agent’s mental states should be considered beyond the scope of this paper.¹ Within this meaning I shall argue that mind-enhancing drugs do not limit or degrade the praise that someone under their influence deserves for accomplishments enabled or facilitated by their use. Roughly, this conclusion is based on the fact that any mental labor involved in the drug-enhanced accomplishments employs the agent’s preexisting mental characteristics, qualities, and knowledge as the foundation to build upon in order to achieve her intellectual accomplishments. Since it is the agent’s personal labor which produces the accomplishments, that agent is deserving of “hard work credit” as well as ensuring her responsibility for future mental states which evolve via this mental labor. The fact that the accomplishments are dependent on the agent’s preexisting mental states clearly links the value of these

1 I am assuming that drugs change the chemical states of one’s brain, which correspondingly affects one’s mental states.

accomplishments with the mind and creditable thoughts of the agent. Hence, an agent satisfying this condition should rightly be considered praiseworthy for her accomplishments. But we must clarify some terms, concepts, and issues that are important to this discussion before we accept this conclusion.

Preliminaries

Throughout this discussion one should assume that if the agent had accomplished the same things but without the influence of mind-enhancing drugs she, and her accomplishments, would be considered praiseworthy and creditable. This means that all her acts are intentional, commendably motivated, the achievements are valuable, and the agent is praised as the ‘achiever of the accomplishments’ not the ‘facilitator’ or ‘vehicle’ of the accomplishments. In other words, the fact that the agent was under the influence of mind-enhancing drugs is the only condition which could make her non-praiseworthy for her accomplishments. Since we are dealing with *mind*-enhancing drugs, one should assume any examples used involve *mental* accomplishments, and that the drug directly affects the outcome or achievement of these accomplishments.

One should also consider that the mental accomplishments being discussed are of value and use to society. Intellectual achievements such as scientific research, philosophical inquiry, literary and poetic writing, business planning, or even creating artwork are the sorts of accomplishments that should be considered in this discussion. Simply taking enhancement drugs to better achieve totally trivial mental goals *may* be permissible, but is probably not praiseworthy. So to be charitable, one should consider the accomplishments discussed throughout as being clearly worthwhile, valuable, and what reasonable people would generally consider praiseworthy.

The Effects of Mind-Enhancing Drugs

The specific effects of the mind-enhancing drugs that will be discussed throughout this paper are also particularly important. Although this is largely a pragmatic issue which is heavily dependent on the specific drug being considered, I believe it is enough for our purposes here to delineate the effects of these drugs and classify them into general groups. Any drugs with effects additional to those specified should be considered in a different category than ‘enhancement drugs’; however combinations of

these effects remain relevant to our scope. These effects can roughly be separated into the enhancement of *mental stamina*, *memory*, or *cognitive capacity*.

Enhancing mental stamina involves lengthening the duration that one can maintain a certain level of mental sharpness or augmenting one's motivations for intellectual engagement. This effect is similar to that of coffee or other common stimulants, except the enhancement drugs being considered are presumably more potent and long-lasting.² To *enhance memory* is to enable people to create or recall memories more easily. Lastly, the *enhancement of cognitive capacity* is the most difficult to define, and to some degree is affected by the enhancement of memory and possibly even mental stamina. It would involve things like improving creativity, the ability to mentally connect concepts or memories, increased clarity of thought, and other such things that people generally consider an increase in intelligence, reasoning, and cognitive processes. These aspects do tend to blend into one another, but this is unproblematic because an agent could experience one or all of these effects and still be a perfectly viable subject of arguments to come.

Accomplishments' Independent Worth

Now that the basic effects of the relevant enhancement drugs have been laid out, the nature and intrinsic value of enhanced accomplishments as events separate from their agents must be shown to be unaffected by the agent's use of drugs. I believe this conclusion is fairly straightforward, and any doubt is a result of confusing the evaluation of the agent with the evaluation of the act. To see what I mean, imagine a mathematical theorem which has just been proven that is as clearly intuitively true as 'the sum of the angles of a triangle add up to 180 degrees.' Whether the mathematician who conceptualized and proved this theorem was on drugs or even accidentally scribbled nonsense on a page which happened to be correct is irrelevant to the value of the theorem. Clearly the state of the agent who produces the work cannot affect the value of it when we consider the product's intrinsic value. However, many claim that values of philosophical, literary, and artistic works are different from the value of mathematical ones, and that they are affected by the people who

2 It is important that a drug doesn't *create* motivations for intellectual engagement as opposed to *augmenting* pre-existing ones. I believe adding such motivations to one's mental character may ultimately constitute a changing of one's identity or intelligence by adding completely new personality features instead of *enhancing* features.

experience them.

In the most basic sense, everything of value is influenced by those valuing it since the meaning of “value” necessarily involves the subjective opinions of the people giving the value. For instance, a physicist may value a mathematical theorem very highly and consider it a great addition to human knowledge, but a hermit living on a mountaintop may consider it totally worthless to the world. In this sense the value of everything *is* affected by those evaluating its importance since value is only manifested in people and their subjective opinions will determine that value. However any objection based on this ground alone is pointless. Holding this position would negate the meaning of praiseworthiness as a personal characteristic, since a person’s worthiness of being praised would be based solely on who would be doing the praising. However, in the interest of producing fruitful philosophy it is perfectly reasonable to recognize the intrinsic ‘praiseworthiness of a person’ and ‘value of an object to people,’ and hence invoke some objective sense of praise and value. Essentially, we can say about a piece of art, literature, or philosophy that it has such-and-such value to the world, which we can attribute to it as an intrinsic aspect.

Essentially, we are talking about the objective value of these works simply as ends in themselves, which is something that can be done at the very least on a practical level.³ The important point here is that whether an author is on drugs or some people hate works produced under drugs, the works themselves have a value independent of that. This is important because even if I can prove later that agents are responsible and praiseworthy for their drug-induced accomplishments, if those accomplishments have little or no value in themselves because they were produced by drug-enhancement, any praise for them is worthless.

Creation vs. Enhancement

The distinction between enhancing one’s mental states and creating new ones was touched upon earlier, but should be explored in detail. This issue addresses the important point that manipulating some preexisting thing is fundamentally different than creating a new one. If a cyclist upgrades from terrible to top-of-the-line equipment and thereby vastly

3 I don’t want to get into a discussion of whether an objective or impersonal viewpoint exists or not. It is enough to assume that at least when considering works of philosophy and literature there is some sort of “objective consensus” that people can come to when considering it practically.

improves his times, this is completely different than if his muscles were all surgically replaced with ultra lean and strong substitutes. This forms a direct and illuminating analogy for mind enhancement drugs, since the former simply enhances his performance *capacity* through a simple non-laborious use of new equipment whereas the latter creates a completely new standard of performance that is independent of the cyclist's previous capacities.⁴ Likewise, a person taking an enhancement drug "simply and non-laboriously" increases her capacity for mental performance whereas a "creation" drug would presumably add ideas or beliefs irrespective of a person's initial mental capacity.⁵ Consequently, our praise of the accomplishments of the upgraded cyclist should be, and generally is, much different than that of the synthetic cyclist. Most people should deny much or all praise for the accomplishments of the synthetic cyclist because his accomplishments are not the result of personal labor that employs his preexisting characteristics, qualities, or cycling ability as the foundational basis which he uses to build his success.

His new muscles create a completely new foundation for him to work with, and regardless of his previous muscle state he is now able to achieve a certain level of performance which anyone receiving his surgery could. This created a new level of cycling ability for him instead of increasing his ability to utilize the cycling ability he already had in himself. This shows that enhancement is fundamentally different from creation, and that this difference has a direct relevance to the praise of the agent. Drug-enhanced performance and accomplishments should be praised based on the same justification which excludes praise for any creation-related performance states or accomplishments, which I believe is a conclusion that matches with our intuitions. However, some may note that this leaves the door open for other sorts of enhancement issues besides this intellectual enhancement discussion, but this problem is mostly beyond

4 When I say "simple and non-laborious" enhancement I mean this externally of buying and setting the equipment up. My point is that the cyclist is unchanged, and by simply picking up a different bike and acting in the same way as before he is able to increase his capacity for cycling.

5 I am not even sure if this type of mental "creation" is possible, but nonetheless it is important to discuss its possible existence. This is why I say they "would presumably" have their effects.

the scope of this paper and can be addressed in a footnote.⁶

Since discussion leads to the same conclusion when looking at intelligence-enhancing drugs, we should consider an illuminating example. Imagine there exists a “genius pill” which automatically raises one’s mental abilities, regardless of the preexisting state of the agent, to an extremely high ability level. Everyone who takes this genius pill has their intellect automatically raised to the same set level or capacity. Then consider another pill which boosts the agent’s capacity to utilize her already existent mental states and capacities. The genius pill is an act of mental creation, in which one’s original states are replaced by the new “genius state”. The enhancement pill, on the other hand, does not enact any form of mental creation, but as its name suggests, it enhances the already present capacities. Any drug which has effects like the genius pill, namely raising intelligence to a predetermined level or somehow creating new knowledge, should be treated like the synthetic cyclist and should not be considered eligible for praise.

On a similar note, the effects of enhancement drugs are better characterized by degree than strict categories of creation or enhancement. The danger one runs into is that dramatic increases in one’s reasoning capacity could be extreme enough to make the preexisting mental states practically insignificant. At that point we no longer want to assert that the agent maintains her praiseworthiness for the accomplishments. The point where one’s preexisting mental states become insignificant is the point where enhancement turns into creation. This is because the point at which one’s preexisting mental states become uninvolved in the achievement of a scholarly goal *simply is* the point where new mental states are created. Yet, the enhancement drugs we want to consider do

6 The big difference between enhancement in sports and academics is the goals or ends being accomplished by the enhancement. The rules and accomplishments of sports are conventional and arbitrary, such as the distance of a “free throw” or length of a football field, whereas the value of knowledge and arts is not arbitrarily assigned. They have goals which are intrinsically valuable and which increase in value as we increasingly excel at accomplishing them. Thus we can regulate sports and stop enhancement and still absolutely achieve their purposes of competition and entertainment without putting undue risk onto athletes. Of course, athletes will only be able to accomplish less spectacular physical feats, but the decision to limit the awe and entertainment of the audience in order to protect the athletes themselves is based on utilitarian considerations. For example, steroids are a common physical enhancer, and they have been shown to do significant damage to people’s bodies. So since these drugs harm athletes and are not necessary to maintain the entertainment that sports provide, it is simply better for everyone to agree not to use enhancement drugs in sports. Yet this conclusion almost certainly does not hold for mental enhancement, which could be used to create vast advantages in accomplishing certain valuable goals.

not affect the agent to this degree, thus they do not run into the creation problem.

SECTION II

The Basis for Intellectual Praise

In the cycling discussion above, I concluded that aspects like “personal labor” and “using preexisting abilities as the foundation for achievements” should be the basis for praiseworthiness in certain cases. The central goal of this paper is to discuss the effect that mind-enhancing drugs have upon the praiseworthiness of the agent for her accomplishments. If these are the conditions which determine this praise, the reasons for them being so must be explained. Although the goal of this paper is not to describe all the conditions an agent must meet in order to be held praiseworthy, it is clear that the use of mind-enhancing drugs is directly relevant to some of them. By looking at a general account of the conditions of praiseworthiness I believe we can identify precisely which condition of praiseworthiness that enhancement drugs relate to. Once we have done this, we will be in a much better position to determine whether enhancement drugs have any actual effect on the praiseworthiness of an agent.

Let us consider a general example of an agent who takes a mind-enhancing drug which enables the performance of an otherwise clearly praiseworthy act, which consists of the conceptualizing of good ideas. We can break this act into three parts: X is A *coming to have* certain good and valuable ideas; Y is A having commendable desires to bring about X; and Z is that A's *desires* to do X gave rise to A *doing* X. The type of enhancement drugs we are concerned with only affect X, thus the issue with mind-enhancing drugs rests on the question: how does the process of arriving at praiseworthy ideas affect the praiseworthiness of the agent? Whatever one's theory of praiseworthiness is outside of this question, we can simply assume the agent fulfills all the conditions and focus on the relevant problem. So from now on, when praiseworthiness is discussed it should be assumed that the point being discussed is not the agent's overall praiseworthiness, but the specific condition of praiseworthiness which is potentially affected by enhancement drugs.

The basis for my defense of praiseworthiness for drug-enhanced accomplishments is that an agent's drug-enhanced accomplishments are the result of mental labor that employs her preexisting mental

characteristics, qualities, and knowledge as the fundamental basis used to achieve her intellectual accomplishments. This constitutes the key condition that must be met in order for an agent using mind-enhancing drugs to be praiseworthy for her resulting achievements. More specifically, this is the condition which must be met in order for the agent to have a legitimate process of mentally arriving at her scholarly accomplishments. I believe this is correct because it allows for enhancing only one's utilization of what she already has within her. The point being that enhancing an agent's ability to use her already present intellectual abilities and actualize her potential accomplishments should not diminish that person's praiseworthiness for those accomplishments.

Preexisting Mental States

Let us break down my proposed condition for praiseworthiness into its aspects of "labor" and "preexisting conditions" in order to form a better understanding of this theory, starting with the preexisting mental conditions. The first thing to do is to provide a clear account of what an agent's "preexisting mental conditions" are. Basically, I will assume that a person's mind simply *is* her physical brain states (or at least is highly affected and correlated to the physical states of the brain).⁷ This means there is a physical state or environment which characterizes the mind, and that someone's reasoning ability, memories, experiences, beliefs, ideas, etc., all correspond to some physical configuration of cells in the brain. It is the "mind" constituted by this basic foundational structure of the brain which essentially constitutes the "preexisting mental states" that I am considering. When an agent begins to think or (as I have called it) performs mental labor, that agent is setting her foundational mental structure into motion, and what is produced depends directly on that preexisting structure of the mind.

I think an interesting way to visualize this situation is to consider the mind as analogous to a machine of sorts. Imagine a giant, old-fashioned machine composed of innumerable gears, levers and pulleys, each with its own determined functions, and imagine furthermore that there are many people who operate this machine by moving or pulling these various components. Assume that the machine can make a variety of different

⁷ This is not meant to be a philosophy of mind discussion, but rather a way to give a clear idea of what I mean. Hopefully if people hold other views concerning the mind-body problem they can translate my current discussion into an analogous one which fits their viewpoint.

products, and although the function of each component does not vary, the manner in which the operators use the machine determines what comes out of it. Furthermore, as the machine is used it also evolves by becoming more efficient and being capable of producing better products. When the machine is not being used it still has a basic fundamental structure that retains the *capacity* to produce things. This situation is basically analogous to an agent's preexisting mental states when the mind is considered to be in a static and pre-labor state. When the workers start using the various components in unison, the machine begins to function and make products. This is analogous to an agent using her preexisting mental states as the foundational basis for her mental labor (which just means she thinks, since she *is* her preexisting mental states).

Notice that if we could take a snapshot of the physical structure of a person's mind, we could consider that structure as having the potential to achieve certain other mental states (ideas, beliefs, etc). More importantly, notice that these later mental states depend directly on the beginning "snapshot" structure as well as the type and amount of mental labor performed. My point is that an agent's preexisting mental states have certain semi-deterministic "mental paths" they will follow once one begins to put mental labor into them. For instance, if enough serious and focused mental labor is "put into" a brain like Einstein's, the potential value and content of the mental achievements that can be realized is different from that of some other brain. Based on the structure of Einstein's brain, he was able to accomplish certain specific feats when he performed mental labor, whereas even when other people work very hard they may simply never be able to arrive at the same ideas. The same goes for any agent's mind, meaning that each of us has a different brain structure and potential for mental achievement. So within the very nature of each of our minds is the potential for certain accomplishments, and depending on the nature of the mental labor performed using this foundational basis, we can achieve these accomplishments. Essentially, we all have unique minds capable of achieving unique accomplishments, and generally whenever an agent's preexisting foundational mental states are the direct basis for the actualizing of an accomplishment, that agent is praiseworthy for it.

This whole discussion may seem very vague and abstract, but I could express this idea fairly simply and still get my point across. Consider that everyone has different interests, ideas, strengths, weaknesses, values, and goals. Based on these different aspects of agents, they are all capable of

producing unique intellectual accomplishments directly suited to their mental nature. Whenever an agent uses her unique mental qualities to achieve some scholarly goal, the fact that the accomplishment of it depends on *her* mental nature makes *her* responsible and praiseworthy for it. So just as before, when an agent employs her mental states as the foundational basis for intellectual accomplishments, she should be praiseworthy for the fact that she achieved them.

In the end, the fact that the accomplishments are dependent on the agent's preexisting mental states clearly links the value of these accomplishments with the mind of the agent, which makes her deserving of praise. Presumably it is the fact that these scholarly accomplishments have qualities of some value and importance, and that these qualities make their agent praiseworthy. Also, it is the mental states and characteristics of the agent which produce these valuable accomplishments, so they are ultimately responsible for the praiseworthiness of the agent. For example, when Einstein formulated his Theory of Relativity, he used his mind to form the concepts and relationships which constitute that theory. It was his thoughts, ideas, and overall content of his mind which produced this important and valuable theory. The fact that Einstein's mind directly relates and connects with his scholarly achievement is what ensures that he is praiseworthy for that achievement. But the mental labor that goes into these achievements is also an important factor for praiseworthiness, and must be explored in detail.

Mental Labor

When a person performs mental labor he is focusing his mind on ideas, relationships, and facts he knows and he is working to produce new ideas, conclusions, or relationships. Mental labor also involves organizing one's thoughts so that others can understand it, coming to understand other people's ideas, and generally an expenditure of mental energy with the purpose of achieving some goal. People often believe their praise for an agent is somehow recognition of that person's effort toward and commitment to her achievement. This is mistaken; mental labor is only indirectly responsible for praiseworthiness because in a practical sense it is a necessary condition for eliciting accomplishments from one's mental states and capacities. In other words, the only *practical* way to produce mental achievements is by performing mental labor (so in some sense it is necessary), but that does not mean it is also necessary for an agent to

deserve praise. However, it is true that over time an agent must be the ‘performer of the mental labor’ in order to be considered responsible for any changed and improved mental states resulting from the labor. Aside from this fact, the mental labor put into a scholarly accomplishment is not what warrants an agent to be praised for it. As we will see, this conclusion will dissolve the problem of enhancement drugs precluding praise, since they only affect an agent’s mental labor; thus their praiseworthiness remains totally unaffected.

Although it seems like mental labor should be a condition for praise, I can show that this is a mistake by looking at a few examples. Imagine an author who writes under some delusional pretenses and performs a great amount of mental labor to produce a book completely different but vastly more valuable than the one he intended to write. It seems clear that the person does not rightly deserve praise *for the value of the book*, which is what we mean when we call an author praiseworthy for his achievements. However, I do believe this author deserves some sort of credit, of a different nature than praise, because he performed the mental labor necessary to achieve this scholarly accomplishment. Perhaps a better way of looking at why praise is, strictly speaking, independent of mental labor is to consider another example.

If someone lacks the condition of having actually performed any mental labor to produce her scholarly achievements, we end up in a very peculiar situation. It is hard to imagine this situation ever actually happening in reality, since we are considering scholarly and intellectual achievements which in a practical sense necessarily involve the use (laboring) of one’s mind. Yet the truth of the matter is that although producing intellectual achievements depends on mental laboring, the praiseworthiness of the agent for those achievements does not. Consider the *Scanner* example, where a machine somehow scans an agent’s brain states and simulates the consciousness of that mind to perform its own “mental labor” to produce some scholarly achievement. Clearly the agent has done no mental labor, but nonetheless he should be praised for the resulting ideas. This might not seem apparent at first, but remember that in a sense it is the *mind of the agent* which is responsible for the achievements, even if the machine performed the labor. Imagine that the machine produces a novel from the agent’s mind which rests completely on the memories, ideas, emotions, and experiences of the agent but which were nonetheless produced by the machine’s labor. It seems clear that the agent deserves the praise for that novel; his mind is totally responsible

for the achievement.⁸ In *Scanner*, the ideas were the direct result of and directly dependent on the agent's mind, hence he is fully praiseworthy for them, even when mental labor is absent.

The last thing to note about mental labor is its importance for the responsibility and praiseworthiness of an agent over time. Clearly the nature of one's mind evolves and changes as he puts thought into various topics. Everyone gains knowledge and perspective, and our mental abilities grow stronger the more we use them, just as our muscles grow stronger with use. The problem then becomes that if mental labor progresses over long periods of time, one's mind is constantly being reshaped and over time the resulting accomplishments are not really the result of preexisting mental states anymore. They are, rather, a product of the *new and evolving mental states* produced by prolonged mental activity. For unaided people this is unproblematic because they are clearly the responsible performers of their mental labor. So they retain complete responsibility for each newly evolved mental state to the same degree as they had for the old mental states that gave rise to the new ones.

However in the *Scanner* example, that agent is not responsible for the mental labor which produces any newly formed "mental states" in the machine. Hence, if the machine were to carry on enough mental labor over time, the agent in *Scanner* would actually lose praiseworthiness for any subsequently resulting achievements, *because they no longer flow from and depend on his preexisting mental states*. Yet within our topic of enhancement drug use, there is no doubt that every agent is the performer of mental labor in achieving her scholarly goals. As long as the agent truly is the performer of mental labor, as opposed to a *Scanner*-type scenario, that agent maintains her responsibility for that mental labor and any newly formed mental states. In order for an agent to lose this responsibility, she would have to somehow lose her identity connection with her labor. But this is simply not a possible consequence of taking the

8 It does not matter that "if it weren't for the machine it would have never happened" or "anyone could have plugged her brain in" because the fact of the matter is that the agent's mind actually did produce these results. Perhaps the agent could have achieved the same things without the help of the machine, so this alone does not preclude praise either. People might make similar claims about books and articles written by other people, but it does not matter if someone else could have written the same material because they didn't and the author did. The fact that someone else could have produced the same result does not affect the original agent's praiseworthiness.

type of enhancement drugs being considered in this paper.⁹

Essentially, enhancement drugs do not make agents lose “responsibility for” or “an identity connection with” their mental labor, and drugs do not make agents lose “responsibility for” or “an identity connection with” their newly formed mental states. Therefore, when agents use enhancement drugs over time, since they must be the performers of the labor, they retain complete responsibility and praiseworthiness for their new mental states, mental labor, and scholarly accomplishments that follow from both. Yet some might argue that although an agent retains responsibility for her labor and newly improved mind, they do not deserve some sort of *hard work credit* for achieving those goals because enhancement drugs were used.

Praise vs. Hard Work Credit

At this point it seems that mental labor deserves some sort of *hard work credit*, whereas the agent deserves *praise* for the actual accomplishment because of the fact that it came from her (preexisting) mental states. Hard work credit should be considered something like the recognition of an agent’s effort toward and commitment to the achievement of a scholarly goal, whereas praise is a positive judgment of the agent due to the value of her accomplishments. It is understandable that these two types of credit have not ordinarily been considered as separate judgments of an agent. This is because they are so closely related and it takes an uncommon situation like enhancement drugs to clearly separate them. However it seems clear, when we consider situations like *Scanner*, that *credit for doing mental labor* and *the mental states that produce scholarly achievements* can be separated, and that the latter is responsible for praise. To relate the *Scanner* example to enhancement drugs, consider that when an agent takes drugs he is enhancing his ability to perform mental labor, *not his preexisting mental states*. In other words, enhancement drugs enable people to achieve a higher quality of mental labor. This mental labor is performed using the preexisting mental states as a base, and having this newfound and enhanced mental ability cannot change the fact that all the labor and subsequent accomplishments are fundamentally based

9 Certain circumstances such as epileptic seizures, extreme drunkenness, or influence of powerful psychotic drugs do break the responsibility and relevant identity connection between the agent and act. However, it is certainly clear that the mind-enhancing drugs being considered here do not have these sorts of effects. If they did, I would not consider them “enhancement” drugs, but rather some other classification.

upon the mental states existent before any drug was used. Thus even if enhancement drugs cause one to lose all the hard work credit for any accomplishments, his praiseworthiness should be unaffected.

Yet it seems clear from the earlier discussion of mental labor that an agent using enhancement drugs retains *responsibility* for labor, which suggests she also retains the hard work credit for it. When on drugs, an agent still has to put forth the mental effort and commitment to achieving some scholarly goal. It is not as if taking an enhancement drug makes mental labor unnecessary in the accomplishments of goals. So it seems that we should give intellectually enhanced agents hard work credit for their accomplishments. The confusion over this deduction arises when we compare a drug-aided agent to an unaided agent. Clearly if they both achieve the same goal the unaided agent might deserve more hard work credit, since he presumably needed to put in more effort and commitment. However, the fact that an aided person deserves less hard work credit does not mean he is any less responsible for his mental labor or newly formed mental states; simply that they needed less hard work. Hence, there is a difference between hard work credit and praise, but this does not affect an agent's responsibility and degree of praiseworthiness for her accomplishments. Furthermore, when we consider an agent's responsibility and praiseworthiness over time, we find that the distinction between praise and hard work credit has no effect on an agent's responsibility and identity connection with her mental labor or mental states.

Lack of Relevant Difference

Finally, we should consider a last example which not only bolsters the concept of preexisting mental states, but also provides a clear argument for why preexisting mental states should determine the praiseworthiness of the agent. Imagine a case involving two otherwise identical people whose only difference is that one is sleep deprived and the other got plenty of sleep; let's call them Tired and Wired. Wired has more mental stamina, can form and recall memories better, and generally can reach a higher state of mental function than Tired. Both are working from the same preexisting mental states, but Wired is simply able to achieve better scholarly goals with less time and effort. However, we would clearly not want to preclude Wired from being praiseworthy for his accomplishments simply because if he lacked sleep (thereby becoming

identical with Tired) he would not have been able to do so. Although as was addressed before, Wired would deserve less hard work credit than Tired if they both achieved the same goal. But again, this is not anything unusual, and certainly does not preclude any of Wired's responsibility and praiseworthiness for his achievements.

This 'Tired vs. Wired' case seems to be a clear analogy to 'enhancement drug use vs. no enhancement drug use', and in fact there is no morally relevant difference between the two cases. In other words, taking drugs or being rested are not in themselves morally relevant facts, but rather the states they produce may be morally relevant. The unaided and Tired agents both simply lack a capacity to utilize their preexisting mental states, and the drug-aided and Wired agents simply have that capacity. According to my arguments, this difference in capacity is simply not morally relevant to the praiseworthiness of an agent. Regardless, if we want to limit praise for an agent aided by enhancement drugs because he has a condition which gives him greater mental capacity, then we must similarly limit the praise of Wired. This is clearly false, so, if enhancement drugs do limit an agent's praise, a relevant and convincing reason must be provided to justify that conclusion.

SECTION III

Arguments Against Praiseworthiness of Enhanced Accomplishments

There are many arguments against holding a mentally-enhanced agent praiseworthy for his accomplishments, but I believe none of them can effectively explain why enhancement negates or limits praise. It is useful to explore these different objections with the help of a hypothetical example. Suppose agents Aided and Unaided are the same person in two identical situations, with the only difference being that Aided is under the influence of mind-enhancing drugs. In this scenario, both agents work toward and accomplish a scholarly goal G. Now imagine there are two different outcomes which could result. In the Efficiency outcome, both produce exactly the same scholarly work which achieves G, but Aided does so with significantly less time and effort expended. In the Quality outcome, Aided produces an outcome with a significantly higher quality than Unaided is capable of producing with any reasonable amount of time and effort. Of course we could also consider the likely case that Aided produces the better outcome with less time and effort, but I think it is more illuminating to consider the implications of the Efficiency and

Quality outcomes separately.

The first argument against praiseworthiness for mentally-enhanced accomplishments is that taking pills is an unnatural form of increasing one's intelligence; thus the agent does not deserve praise. I believe it is a weak interpretation of this objection if we assume it to mean that drugs are all artificial and people should lead a "natural" lifestyle. This type of argument has been employed in many religious arguments and I believe it simply fails to provide any adequate grounds for objecting. However, it is interesting to interpret this argument as meaning people should only be praiseworthy for accomplishments achieved by abilities they are born with or develop through non-artificial means, such as unaided reading, studying, and thinking. So the Efficiency outcome is much less praiseworthy, since Aided unnaturally evaded the usually necessary time and effort to achieve the goal. Similarly, in the Quality outcome Aided exceeded the natural limit of his ability. I believe the naturalness argument stems from a basic belief that people should be praised for what they are able to accomplish without aid, and when one starts getting help by using "artificial" means to produce extra abilities, those additions deserve the credit since the agent is not achieving his accomplishments on his own anymore. While this idea of "doing it on one's own" has some merit, it is almost impossible to separate natural means of gaining abilities from artificial ones, as well as when an agent loses praise because of the help he receives. For instance, are herbal supplements and caffeine natural? The answer does not seem clear, and the question most likely is not resolvable. Also, is having a great idea that was only conceived because one studied another person's work getting too much help? It seems clearly not, yet it also seems that the help was extremely determinate of the outcome. In the end, the argument against drugs because they are artificial and therefore degrade praise is not convincing.

A similar argument states that it is unfair to praise Aided more than Unaided, or perhaps even at all, for producing better and more efficient accomplishments simply by taking a pill. This argument seems to hinge on the fact that presumably anyone can effortlessly take a drug and thereby intellectually outperform an equivalent person, which is unfair. However, I simply do not think fairness is relevant to enhancement drugs, because anyone presumably can take them and they still require considerable effort and knowledge in order to accomplish anything of worth. Let us assume that Einstein's brain was just more capable of thought due to its genetic and physiological makeup. Yet he had to put in massive amounts of time

and effort to build upon those foundations to produce his wonderful ideas. Through no intention, effort, or responsibility of his own he was far more capable of achieving important scholarly goals than almost any other person. We do not even consider limiting his praiseworthiness for these accomplishments just because his capacity for reasoning was higher than most people's. Similarly, enhancement drugs increase the user's capacity for reasoning through no effort or responsibility, so it seems inconsistent to praise one but deny it to the other.¹⁰

The next argument is that people taking enhancement drugs somehow lose their identity, and thus they have no agent responsibility, and therefore praise, for accomplishments achieved using these drugs. This question relies on the definition of an agent and whether or not the effect of enhancement drugs makes a person no longer fulfill the conditions of being the agent of her accomplishments. Clearly the person retains her identity while under these drugs, so we know she is the same person who performs the accomplishment. But perhaps the agent-act relationship is disrupted while leaving the person-act relationship intact. However, I believe this does not happen since the actor performs the mental labor and does so using the knowledge and mental characteristics unique to himself. Such a proximity to an intimacy between the actor's nature and the nature of the act makes his status as an agent legitimate. Whatever one defines the agent of an act as, it should include a person who intentionally carries out the work necessary to produce the act and furthermore that the details of that act are determined by the person's characteristics. Thus taking mind-enhancing drugs does not negate one's identity or agency for accomplishments.

In the end, none of these arguments is convincing. I believe that mind-enhancing drugs simply do not limit the praise that an agent deserves for her mental accomplishments while taking them.

SECTION VI

Conclusion

This paper was meant to show that when an agent takes intelligence-enhancing drugs, that fact does not affect his praiseworthiness for any valuable scholarly goals he is able to achieve because of them. This is true

¹⁰ Perhaps people using enhancement drugs deserve even more praise because unlike Einstein they may have intended to increase their cognitive capacity in order to produce helpful scholarly accomplishments. This certainly sounds praiseworthy to me.

because the agent's drug-enhanced accomplishments are the result of mental labor that employs her preexisting mental characteristics, qualities, and knowledge as the fundamental basis used to achieve her intellectual accomplishments. The fact that it is the agent's own personal labor which produces the accomplishments makes the agent deserving of hard work credit as well as ensures her responsibility for future mental states which evolve via this mental labor. The fact that the accomplishments are dependent on the agent's preexisting mental states clearly links the value of these accomplishments with the mind of the agent, which makes her deserving of praise. Thus people deserve praise for their drug-enhanced accomplishments.

However, it is extremely important to realize what this paper did not and was not trying to prove; namely that it should be morally permissible for anyone to use mind-enhancing drugs in order to achieve their goals. There are numerous arguments which suggest that it may not be morally permissible to use enhancement drugs in some circumstances. Furthermore, it seems that there are cases where enhancement drugs should be prohibited as a matter of convention, such as when people take standardized tests. This topic of permissibility is in itself a serious and complicated issue, and one which simply was not discussed here.

Special Thanks

I would like to specially thank Professor Holly Smith for her comprehensive and insightful comments, not only on drafts of this paper but also in her lectures, which encouraged thought into the issues I have written about.

Hobbes and the Limits of Political Obligation

Konstancja Duff
University of Cambridge

Why should we obey the government, and when, if ever, do we have the right not to?

Hobbes's absolutist response to this question is as powerful as it is disturbing: rational self-interest, he argues, demands total submission to the state. By focussing on his account of the necessary conditions for just resistance, my paper aims to explore the cogency of the idea of utter subjection for our own good. I argue that even if we go along with everything Hobbes says about human nature and the construction of political obligation, closer examination of some of the epistemic concerns inherent in his account suggests that we are not compelled to accept his radically illiberal conclusions.

Why should we obey the government, and when, if ever, do we have the right not to? This question, essentially one of how the justification of political obligation can ever be consistent with its limitation, has framed political discourse for millennia. Never did more hang on the answer than when, as Parliament executed the King along with his divine right to rule, and the explosion of print culture swamped the streets with rhetoric, Hobbes published his seminal defence of absolutism: *Leviathan*. Arguing from a bleak vision of the natural condition of mankind to the rational necessity of renouncing our rights and investing them in an all-powerful sovereign, he concludes that our obligation to obey such a sovereign is absolute, right up to the moment when he actually attempts our destruction. Although unmistakably defined by the events of his time, Hobbes's analyses of human nature, rational self-interest and the politics of power are disturbingly relevant to our own era. By focusing on his account of the necessary conditions for just resistance, this paper aims to explore the cogency of the idea of utter subjection for our own good. Ultimately, I shall argue that even if we go along with everything Hobbes says about human nature and the construction of political obligation, closer examination of some of the epistemic concerns inherent in his account suggests that we are not compelled to accept his radically illiberal conclusions.

In order to appreciate the subtleties of Hobbes's position on the conditions for just rebellion, it is necessary to trace its roots in the narrative he presents concerning the state of nature and the generation of a common-wealth. Painting an apocalyptic picture of man in his natural state – a state of perpetual 'warre, as is of every man, against every man... And the life of man, solitary, poore, nasty, brutish, and short'¹ – Hobbes argues that this follows inevitably from certain essential features of human nature. People, he observes, are by nature effectively equal; the weakest can kill the strongest, and since none have sufficient strength for security, no stable hierarchy can develop. From this basic equality flow three primary sources of conflict: competition for resources; the desire for glory and status; and, most importantly, *diffidence or fear*. Diffidence as a motivating factor is to be understood chiefly as embodying the idea of *rational anticipation*; the accompanying notion of some kind of epistemic calculus of risk is thus brought into play. The effect of such a calculus on

1 Hobbes, *Leviathan*, ed. R. Tuck, rev. edn (Cambridge: CUP, 1996), pp. 88-89. All subsequent references are to this edition.

the actions of self-preserving agents is such that, even if only a few are motivated to grab all the resources and power that they can, none can be secure in their more modest portion, and so must strive to augment their power merely to survive. Crucially, for Hobbes, this war of all against all is not merely part of some historical narrative; it represents a real and present danger from which, as we shall see, we can save ourselves only by total subjection to dictatorship.

A further factor driving the conflict, and one which is ultimately key to Hobbes's treatment of the subject, is the unique human capacity for language and consequently for the abuse of language. We can, as Hobbes puts it, 'represent to others, that which is Good, in the likeness of Evill; and Evill in the likeness of Good' (*Lev.* pp. 119). This rather neatly encapsulates the fear that drove the intense contemporary debate surrounding the dangers of language, and specifically of rhetoric. As Hobbes (among others) recognised, though, it is but one symptom of the perennially troublesome mediation that language provides between us and any external reality.

Irredeemably tied up with the problem of language is the well documented fact that people's individual judgements just do not concur. For not only does language bring with it the possibility of misrepresenting reality to others; it also allows that there be an immeasurable number of different ways of representing reality to *ourselves*. We can see this connection most clearly spelled out in the *Elements of Law*:

In the state of nature, where every man is his own judge, and differeth from others concerning the names and appellations of things, and from those differences arise quarrels, and breach of peace; it was necessary there should be a common measure of all things that might fall in controversy...²

Crucially, in the absence of such a common measure, when each person is judge of the justness of her own fears, and 'private Appetite is the measure of Good, and Evill' (*Lev.* pp. 111), there can be no binding laws or covenants.

In the state of nature there is but one fundamental law: the right of self-preservation. Hobbes endeavours to show that the whole structure of political obligation (including its limits) can be derived from this self-evident principle. Working on the grounds that security

² Pt.II.ch.10, 8, in T. Sorell (ed.) *The Cambridge Companion to Hobbes* (Cambridge: CUP, 1996), pp. 185-86.

is impossible in a state of war, and consequently reason dictates that ‘every man, ought to endeavour Peace, as farre as he has hope of obtaining it’ (*Lev.* pp. 92), Hobbes proposes a set of ‘Lawes of Nature’ conducive to peace. It is interesting to note that this ‘morality’ has the essential form of a hypothetical imperative: *if* you desire your own preservation, act according to these rules. So the normative force of his law is contingent on a desire; nevertheless, it is a desire which universally obtains.

There is, of course, another important sense in which Hobbes’s Laws are hypothetical in structure. The obligation to act according to them will hold only if one is *secure in the belief that others will do likewise*, this proviso being a direct consequence of the Laws’ derivation from a calculus of self-interest:

...if other men will not lay down their Right, as well as he; then there is no Reason for any one, to devest himselfe of his: For that were to expose himselfe to Prey, (which no man is bound to) rather than to dispose himselfe to Peace. (*Lev.* pp. 92)

We are now in a position to see the full significance of the right to private judgement which exists in a state of nature: if each person need only be bound by their covenants if they judge that it is safe to do so, and everyone knows that everyone else is performing this same calculation but without any objective standards of judgement, then a rational agent will never achieve the security required for obligation. The question of what we can *rationaly anticipate* is again paramount.

In addressing this problem, there is a crucial distinction Hobbes draws between being bound *in foro externo*, that is, bound to *act* according to a law, and being bound *in foro interno*, that is, being bound only to the *will* – the ‘unfeigned and constant endeavour’ (*Lev.* pp. 110) – to act according to it. In a state of nature, the Laws can oblige only *in foro interno*, for to be obliged *in foro externo* would contradict the very basis of those laws. Cooperating without assurance that others will do likewise (as being bound *in foro externo* would require) is distinctly contrary to rational self-interest. In order to get to a situation where the Laws of Nature *can* bind in this latter sense, we must make the transition from state of nature to civil society, and recognising precisely what this move involves is the key to Hobbes’s stance on the necessary conditions for just resistance.

In a civil state, we can be bound *in foro externo* because what we can reasonably anticipate has changed: the threat of the sovereign’s sword

makes it prudent for each self-preserving agent to cooperate, so we can reckon on their acting accordingly. The tricky bit, of course, is how we can covenant to institute this power without the power already being in place to make such a covenant binding. What we require at this point, in order for it to be rational for us to cooperate, is reasonable certainty that others will cooperate also, and in this sense Hobbes's model prompts obvious comparisons with game theory's Prisoner's Dilemma. Both represent coordination problems in which self-interested agents, because of the nature of the set-up, are rationally compelled to make choices which lead to a (self-interestedly) sub-optimal result.

However, there is a crucial difference between the scenarios, which Hobbes exploits in order to escape the trap of defection: the Hobbesian agent is indeed self-interested, but her primary goal is survival, not optimisation of advantage, and this has important implications for her actions³. Being bound *in foro interno* to the Laws of Nature, in the scenario where all are agreeing to transfer their rights to the would-be sovereign, she will be highly motivated to avoid the outcome where she renounces her rights but others do not renounce theirs, but far less motivated to engineer the (in some sense optimal) situation where others renounce their rights but she retains hers. The factor which tips the balance in favour of cooperation is the fact that the latter outcome would be one in which her life was still in danger, for in revealing herself as a deal-breaker she would have made herself the enemy of those who had cooperated, and their newly empowered sovereign would have reason to destroy her. Hobbes, therefore, is able to present a coherent picture of how the Laws of Nature could come to be binding *in foro externo*.

Having put in place the ground-work on the basis of which Hobbes's position on the limits of political obligation is to be understood and assessed, let us now turn to the position itself. Essentially, Hobbes's line seems to be that a person is bound *in foro externo* to obey the sovereign's commands up to the point where the sovereign's sword is, literally or metaphorically, at her throat. The question of precisely *how* literally or metaphorically shall be considered later. For now, it is necessary only to draw attention to fact that, for Hobbes, 'the motive, and end for which this renouncing and transferring of Right is introduced, is nothing else but the security of a man's person' (*Lev.* pp. 93). The covenant to obey the sovereign, though prior to (and necessary for) almost all other

3 Harrison, R., *Hobbes, Locke, and Confusion's Masterpiece: An Examination of Seventeenth-century Political Philosophy* (Cambridge: CUP, 2003).

commitments, cannot be prior to the commitment to self-preservation which motivated that covenant in the first place.

Given that an account of the rationality of subjection is being given purely in terms of self-interest, the most obvious worry is that the Hobbesian system might collapse into straight-forward egoism, much as some have argued that rule-consequentialism collapses into act-consequentialism. Hobbes confronts this objection head-on in his reply to the hypothetical Foole, who ‘questioneth, whether Injustice [that is, disobedience]...may not sometimes stand with that Reason, which dictateth to every man his own good’ (*Lev.* pp. 101). The reply focuses on the overriding importance, when it comes to judging the rationality of an action, of the epistemological question of what you can know in advance or *reasonably predict*.

...when a man doth a thing, which notwithstanding any thing can be forseen, and reckoned on, tendeth to his own destruction, howsoever some accident he could not expect, arriving, may turne it to his benefit; yet such events do not make it reasonably or wisely done. (*Lev.* pp. 102)

Incidentally, this may serve as a warning to those who would suggest that the successful liberalisation of society that has taken place over the past few centuries can straightforwardly disprove Hobbes’s thesis.

For the next stage in my argument, I take as a basis the classic case in which it might be thought that we have the right to resist the government: when it is no longer acting in the interests of the people. In examining the various ways in which the things Hobbes says may be brought to bear on this matter, it should be possible to draw together the different strands of his position, and to see precisely where the tensions lie. The principal question that Hobbes would pose to someone who asserted that we can resist the sovereign when they are acting against the interests of the people would be: ‘Judged by whom?’⁴ The sovereign, as law-enforcer and arbitrator, has the power to decide all controversies,

⁴ The reply given by many political thinkers, including Locke, would be that ‘the people’, taken as a whole, or their representatives, can judge the sovereign’s actions. Hobbes is determined to rule out this possibility, and in Chap. XVI of *Of Persons, Authors, and things Personated*, he argues: (a) that ‘...it is the *Unity* of the Representor, not the *Unity* of the Represented, that maketh the Person *One*... And *Unity*, cannot otherwise be understood in Multitude’; (b) that the Sovereign is the one and only representative of the people; and consequently (c) that the idea of a judgement by ‘the people’ is an incoherent notion. Thus we are left with private judgement, upon which I shall focus.

and incorporated within this is the power to judge what is in the people's interests. In our original covenant, we submitted '[our] Wills, every one to his Will, and [our] Judgements, to his Judgment' (*Lev.* pp. 120). Private judgement, for Hobbes, was the thorn in the side of peace, and was rationally renounced in the interests of security. And, as the reply to the Foole makes clear, we cannot just take back our rights when it suits us. In the case of the right to judgement, it is still more absurd to suggest that we can take it back when it suits us, for how can we judge when that is the case? There appears to be a dilemma: either we just can't tell what our own interests are, or we never really gave up our judgement in the first place.

This is where plausibility becomes a problem for Hobbes, for he must endorse one of two interpretations: (a) that our submission to the sovereign entails that we literally *believe* all his judgements to be correct; or (b) that we may have our private judgements, but never act upon them when they contradict the judgement of the sovereign. Both of these positions are problematic. The first seems to fall into the same difficulties as Pascal's Wager, i.e. that *beliefs just don't work like that*. We cannot generally believe something merely because we have reason to think that such a belief will be in our interests. Or even if such a process of self-brainwashing were possible, its results would surely not be stable enough to rely upon as an integral element of civil obedience. After all, it is an important methodological feature of the Hobbesian approach that we base our political system only on those aspects of human nature which we can reliably predict will obtain in a critical number of cases.

It seems far more likely that Hobbes would lean towards the second reading: certainly the idea of a judgement that we don't, for whatever reason, act upon does not immediately strike one as incoherent. We can think of normative judgements as being intrinsically motivating but with an implicit *ceteris paribus* clause, which allows for there being situations in which we don't act on them. There is, though, something very odd about the idea of entirely divorcing our judgements about what is right (or even about what is in our interests) from our judgements about what we should do. If only the latter are prohibited by political obligation, though, then it seems that this is what the Hobbesian position demands us to do. To put it another way, the Hobbesian world of political obligation is one in which we must recognise that, for all our private normative judgements, the *ceteris paribus* clause might well never obtain because the judgement of the sovereign is always the trump card. *Quite*

what would it mean to have a private morality in that context is open to question. Of course, Hobbes's radical view that there is no such thing as injustice apart from disobedience to a positive law might allow him to do away with such a morality altogether⁵. Part of the attraction of his account, though, seemed to be that it obtained *regardless of*, indeed partly *because of* the empirical fact that our private moralities conflict. If his version of political obligation turns out to preclude any substantial normative commitments beyond the commitment to obey the sovereign, his solution begins to look, ironically enough, a bit unrealistic⁶.

Still relating to the idea of renouncing our right to judgement, there is a further worry concerning the internal cogency of Hobbes's position. This should become clear when we consider the one situation in which Hobbes *does* allow for disobedience: when there is a clear and present threat to one's life. Hobbes's own premise demands that he recognise this as an exception. Yet doing so brings with it an important readjustment of precisely what is allowed to go on in the head of a subject. For once we have recognised even the *possibility* of ever justly acting on our private judgements contrary to the will of the sovereign, there arises a need for some kind of 'meta-judgement' to decide whether any given situation is one of these exceptions which Hobbes admits.

But, the Hobbesian could argue, this thought is based on a confusion, and indeed an over-intellectualisation of the issue. The only 'meta-judgement' which need come into play is the entirely unproblematic one of realising the obvious fact that you are being attacked and the need to react accordingly. It need not have consequences beyond the limited realm of its application. This reply, though, can be seen to be specious when we consider the practical situations in which Hobbes himself would admit the practical judgement that one's life is in danger. No one can be obliged, he says, to bear witness against himself (where there is no offer of pardon), for in doing so he would be bringing about his own

5 'Where there is no common Power, there is no Law: where no Law, no Injustice' (*Lev.* pp. 90). Hobbes does, however, say that there can be 'iniquity' in such circumstances, which implies that he is not a *complete* moral constructivist: he is allowing for the existence of *some* moral standpoint which does not require 'the Sword' for its legitimacy.

6 This would depend on just what 'having a normative commitment' consisted in. In arguing against religious commitment ever taking precedence over commitment to the sovereign, Hobbes again puts it in terms of the relative certainty of outcomes: 'there is no naturall knowledge of man's estate after death' (*Lev.* pp. 103). It is not clear how this would apply to normative commitments which were not self-interested in even this broader sense; Hobbes might well deny their existence.

death or imprisonment⁷, and this is contrary to the whole purpose of political obligation. It would clearly be ridiculous, then, to suggest that the judgement that one's life is in danger amounts to nothing more than 'I can see his sword coming at me, so I must be being attacked'. In fact, once we start envisaging scenarios, it rapidly becomes apparent that there is no sharply delineated set of events which fall under the heading: 'attacks on my life'. Why should I be permitted to judge that I am under attack when the King's men come knocking on my door, but not when I hear that they have set out towards my house? Or when I hear that the King has written an arbitrary law under which I can be arrested? It seems much more plausible to say that these events lie on some kind of continuum; consequently, we must justify any decision we make about where to draw the line, or whether to regard political obligation itself as lying on a comparable continuum. Hobbes's position, read as prohibiting all but the most desperate resistance, can be seen to rest on the dubious assumption that it is psychologically and rationally possible for me to deliberately refrain from realising that my life is in danger until the fact (or the sword) is staring me in the face. Or if I am allowed to realise, but not to act, Hobbes's insistence that the people have surrendered their judgement again seems on shaky ground, and a central constituent of his ideal of absolutism is undermined.

We have now identified two interconnected factors which are crucial to assessing the permissibility of resistance in any given situation: first, the intelligibility of our obligation to suspend private judgement; second, the probability of death in not resisting. Rather than being completely polar, these elements are better understood as varying on some kind of sliding scale. In illustrating his position, Hobbes tends to select scenarios in which both elements are, so to speak, turned up to the max – there is 'certain and present death in not resisting', and the suggestion that we should suspend our private judgement is nonsensical. To put it another way, our normally overriding obligation to obey the sovereign is downgraded to an *in foro interno* duty because all the red lights on

7 At all the crucial points, Hobbes emphasises the primacy of self-preservation, but in saying that we can resist the sovereign to avoid imprisonment, he seems to imply that there are some lives that we might rationally prefer to risk death than endure. He doesn't make this strand of his thinking explicit, but it lies rather uncomfortably with the rest of his case. After all, if we can resist imprisonment, why can't we resist disastrous economic policies or tyrannical legislation which causes us comparable misery? Hobbes's response must be that we have renounced our right to judge the probable consequences of the sovereign's actions; this is the claim that my main argument is intended to challenge.

our self-preservation alarm are flashing, and we cannot be expected not to act to defend ourselves. Having recognised the mechanism at work, though, there arises the definite suspicion that Hobbes really hasn't justified drawing the line where he does. Can he really assume that it is possible for us to suspend or disregard our private judgement right up to the point when we are being physically attacked? Is it really plausible that we should not be allowed to realise our peril in any less obvious situation, and if we do realise, that it should not be reasonable for us to act? My contention is that, to these questions, Hobbes gives us no satisfactory answers.

To conclude, then, examination of the structure of Hobbes's argument for political obligation brings to light two internal tensions which undermine his conclusions on the matter of just rebellion. It is interesting to note that both these concerns are essentially epistemological, and make sense when we consider the fact that his system has its foundations in the idea of *rational prediction*. By demonstrating that Hobbes's judgement as to where the limits of political obligation must be drawn depends on some crucial epistemological assumptions which we need not go along with, I have attempted to show that there may be more scope for just resistance even within the Hobbesian system than Hobbes is willing to admit.

References

- Hobbes, T., *Leviathan*, ed. R. Tuck, rev. ed. (Cambridge: CUP, 1996)
- Hoekstra, K., 'Hobbes on the Natural Condition of Mankind' in P. Springborg (ed.) *The Cambridge Companion to Hobbes's Leviathan*, (New York: CUP, 2007)
- Harrison, R., *Hobbes, Locke, and Confusion's Masterpiece: An Examination of Seventeenth-century Political Philosophy*, (Cambridge: CUP, 2003)
- Ryan, A., 'Hobbes's Political Philosophy' in T. Sorell (ed.) *The Cambridge Companion to Hobbes*, (Cambridge: CUP, 1996)
- Skinner, Q., 'Hobbes on Persons, Authors and Representatives' in P. Springborg (ed.) *The Cambridge Companion to Hobbes's Leviathan*, (New York: CUP, 2007)

Wittgenstein on Scepticism in “On Certainty”

William Crouch
University of Cambridge

In this essay I look at Wittgenstein’s response to scepticism in *On Certainty*. I discuss McGinn’s non-factual account and Wright’s epistemic account. I argue that the former cannot but that the latter can provide a satisfying response to the sceptic, one which both appreciates the truth in scepticism (unlike Moore) and provides reason why we should be content to continue our everyday practices.

Moore's Proof

On *Certainty's* main focus is Moore's response to scepticism in 'A Defence of Common Sense' and 'Proof of an External World.' The classic example argument that is given in these articles comes from the latter and runs:

(A)

1. Here is one hand, and here is another.
2. Therefore (from (1)), two physical objects exist.
3. Therefore (from (2)), an external world exists.

As we can see, Moore's is an unqualified defence of common sense. His argument has been a source of puzzlement for philosophers, in light of the fact that it appears for all the world to flamboyantly beg the question; Moore appears to be acting as if he has had a philosophical lobotomy.

Wittgenstein, like many other commentators, thinks that Moore's proof does not work. Nonetheless, the propositions which one holds with absolute certainty, examples of which such as (1) Moore gives in his articles, *do* form an interesting class for Wittgenstein. It is this class of propositions that is the focus of his thought in the text.

Hinge Propositions: A Schema of Wittgenstein's Response to Scepticism

Wittgenstein labels these propositions 'hinge propositions'. He suggests that these propositions have a special role in our practices of inquiring. The metaphor he uses is that these propositions form the hinges around which our inquiries turn. The suggestion, then, is that these propositions are not aspects within the practice of inquiring, but rather in some way constitute it. They are thus excluded from epistemic evaluation.

Examples of hinge propositions that Wittgenstein gives are: 'I have two hands'; 'The world is more than five minutes old'; 'My name is L.W.'; and ' $12 \times 12 = 144$ '. Four points about them as a class are worth noting (Pritchard, forthcoming, 5):

1. The propositions form a heterogeneous class. There is nothing that apparently links them all apart from the fact that Wittgenstein himself believes them with complete certainty. There is no function for determining the class of hinge propositions.
2. What may act as hinges in one context may act like normal

propositions in another. For example, though in the standard case - and in the context in which Moore asserts it - the proposition 'I have two hands' is held with absolute certainty, such that the speaker does not need to *check* that she does have two hands, this may not occur in another situation. After an operation when the speaker is wrapped in bandages he may very well doubt that proposition, and need to check by looking.

3. They are indubitable, but in a sense different from Cartesian indubitability. That is, they are not always incorrigible or self-evident, like the cogito - this can be seen from the above example. In this way, their indubitability is no indicator in itself of their likelihood of truth.
4. They are groundless: there are no *reasons* that one can give in support of them that give grounds stronger than just asserting the proposition itself. And in ordinary practice we feel no need to provide grounds for them. Note that this only applies to hinge propositions *when acting* as hinge propositions - see case (2) above.

Wittgenstein wants to say that there is something wrong with Moore claiming that he knows these propositions. Likewise with doubt: Wittgenstein correctly judges that both doubt and knowledge claims are illegitimate when applied to hinge propositions (On Certainty, 58).

The big question for commentators is why this is the case. So far, the account of hinge propositions appears to be grist for the sceptic's mill: after all, these hinge propositions are groundless, neither self-evident nor incorrigible and knowledge-claims about them are problematic. The main subject of analysis in this essay, therefore, will be what it is about hinge propositions that renders them immune to sceptical attack.

One quick answer that we should give no credence to is the idea that we may immediately say that the sceptic's doubts are illegitimate because they occur in an illegitimate philosophical context: the words have been robbed of their meaning by being uprooted from their normal usage.

The reason that this quick response should appear doubtful is that there is nothing *peculiar* about the sceptic's questioning apart from its generality. The sceptic is merely making an extension of our natural doubt - seeing that we sometimes doubt the veracity of certain propositions on the basis that we are hallucinating (etc) and asking why we do not *always* raise these doubts. Indeed, the sceptic could be viewed,

as Stroud views her, as adopting a *refined* conception of justification. In normal circumstances we simply do not have time to check that each of our beliefs are justified; but, ideally, we would like to. This was Descartes' motivation in the *Meditations*.

With this quick fix out of the way, let us now look at how commentators have interpreted why it is that hinge propositions are exempt from doubt.

Interpretation of 'Hinges': The Non-Factual Account

McGinn provides an example of this popular interpretation. Hinges are to be regarded as *rules* of enquiry, rather than fact-stating propositions. They help to define a practice: the proposition 'an external world exists' is simply a formulation of a rule that we implicitly learn and must adopt in order for us to take part in the language-game of talking about physical objects. In light of this, they cannot be *wrong*, any more than the rule 'in chess, a bishop can only move along diagonals' can be wrong. They are examples of the way we regulate, rather than products of, enquiry.

There are two salient advantages to this interpretation. Firstly, it gives a good account of why it is that there is nothing that we are ignorant of even though we do not *know* these propositions. If hinge propositions were fact-stating then they *would* appear to be apposite objects of knowledge, and the sceptical argument would go through.

Secondly, McGinn's interpretation does not conflict with the plausible Closure Principle:

Closure Principle: If S knows that p and S competently deduces q from p, then S knows that q.

In contrast, if hinge propositions *were* fact-stating, then Wittgenstein would be committed to a denial of this principle, for he would have to deny that the following argument is valid:

(B)

1. Moore knows that he has two hands.
2. Moore competently deduces, from the fact that he has two hands, that an external world exists.
3. Therefore, Moore knows that an external world exists.

Wittgenstein would have to deny Closure because he would have to say that (1) is true but that (3) is false. The intuitive plausibility of the principle means that this would be detrimental to his account. However, this problem does not arise for Wittgenstein on McGinn's interpretation

because (3) is simply not a fact-stating proposition which can feature as part of an argument in the above form.

The real problem with McGinn's account is that it does not resolve the sceptical worry. This is because, in general, we have rules for a specific purpose. The only case in which rules are not evaluable is when the practice which the rules constitute has no aim or purpose at all. But when there *is* a purpose, we may evaluate the rules in the light of this purpose. And what is a main purpose of scientific inquiry? Precisely that of divining the truth: that is, of accurately representing the world. And it is precisely this that the sceptic is saying that our practices *don't* do. So the sceptical worry may now be reformulated as questioning whether the rules which constitute our fact-stating discourse are really well suited to the purpose of representing reality.

On this account the labeling of Wittgenstein as a linguistic idealist may appear appropriate. For if no sense can be made of the idea of representing reality then the above formulation of the sceptical worry will not be pressing because it was never the business of our language to represent reality in the first place. This is an example of internal realism, as defended by Putnam, and would be the most likely immediate reply to the objection.

However, we cannot be happy with this as a response to scepticism, for two reasons. Firstly, it is of no help simply to deny that we were ever after the goal of representing reality in the first place. It is clear that we *were* always aiming at representing reality – that is, internal realism is clearly revisionary rather than descriptive – because if it were not the case, we would never have found the sceptic to be as disturbing as he is. Rather, he would be on par with a philosopher who denies that we can meaningfully talk about representing reality in our discourse about humour or taste. In light of this, internal realism looks more like an *acceptance* of scepticism rather than a refutation.

Secondly, we have to have grounds for supposing specific propositions to be non-factual. Wright at one point argued that this could be given via the existence of irresolvable disputes. So, for example, there doesn't appear to be anything that can resolve fundamental disagreements over what is funny or over matters of taste. This provides grounds for thinking that these propositions are not in the business of representing an objective reality. The hope would be that the same could be said of hinge propositions: the disagreement between the sceptic and Moore (for example) appears irresolvable, so this gives us grounds for thinking

that hinge propositions are not fact-stating.

However, as Wright now argues, this is not satisfactory when applied to hinge propositions. This is because we need prior grounds for supposing that, if the propositions *were* fact-stating, the fact stated would be detectable. In the case of ‘there exists an external world’ it appears that the disagreement would be irresolvable *whether or not* there was a fact of the matter to which the proposition corresponds. So the irresolvability of the disagreement can’t cast doubt on the factuality of the proposition.

Interpretation of ‘Hinges’: The Epistemic Account

Wright is the main proponent of what I call the epistemic account. He argues that we do *know* hinge propositions – this differentiates him from the previous interpreter – but that they are ungrounded. We thus may not speak as having reasons for them. Wright fleshes this out in terms of the failure of the transmission principle:

Transmission Principle: If S knows that p on the basis of supporting ground G, and S competently deduces q from p (thereby coming to believe that q while retaining her knowledge that p), then G is sufficient to support S’s knowledge that q.

This principle should be contrasted with the closure principle. Whereas the closure principle states that knowledge transmits across the competent deduction of q from p, the transmission principle states that justification as well as knowledge transmits across the competent deduction of p from q. It is thus a more demanding principle than the closure principle.

Wittgenstein, on Wright’s interpretation, accepts the closure principle, and for that reason would accept argument (B) as valid. But he does *not* accept the transmission principle, and so would think that the following argument is not valid:

(C)

1. Moore is justified in believing that he has two hands from his perception of them.
2. Moore competently deduces, from the fact that he has two hands, that an external world exists.
3. Therefore, Moore is justified in believing that an external world exists because of his perception of his hands.

The reason for this is that the defeasible inference from one's sensory impression of a hand to the existence of a hand *presupposes* that an external world exists: the inference is only legitimate given that an external world does exist.

So Moore *is* wrong in saying 'An external world exists'. But this is because he has no grounds for his assertion (hinge propositions are groundless), *not* because he doesn't know it.

As I have phrased it so far, however, this sounds very much like scepticism again. The obviously pressing question, vital to understanding why scepticism is not a problem for our ordinary practices, is why we ought to accept these propositions even when they are groundless.

Wright's reply is that there is an insight in scepticism, which is that it shows the limits of our justification. However, one cannot but take hinges for granted: any attempt to justify them results in cognitive paralysis. The justification that the sceptic demanded was never needed.

Pritchard originally, and naturally, interpreted this as a pragmatic response to the sceptic: we simply *have* to accept these ungrounded propositions, so there is no point in arguing about it (forthcoming, 23). This is obviously deeply unsatisfying as a philosophical response; it certainly would not convince the sceptic. Like internal realism, it appears as if one is backhandedly admitting defeat to the sceptic.

The correct interpretation of Wright's position, rather, can be seen in the following thought experiment. Imagine a potential inquirer, originally in isolation from any practices, who wishes to discover truths about the world. The potential inquirer has only two alternatives: he can either be a sceptic, and not adopt any practices at all; or he can adopt a method of inquiry, and in doing so must adopt certain propositions that, of necessity, will remain groundless during the inquiry. If he takes the former option then he is *guaranteed* not to form any true beliefs about the world. If he takes the latter option then it becomes at least *possible* that he may form true beliefs about the world. In this way we can explain why, as a rational agent, he ought to adopt a method of inquiry.

The correct attitude, then, is one of epistemic responsibility: we don't need evidence for a presupposition of inquiry as long as there is no extant evidence against it.

Conclusion

We have seen, therefore, that there *is* a satisfying response to the sceptic

which can be gleaned from *On Certainty*. This response involves acknowledging the groundlessness of propositions when they are acting as hinges and thus why Moore was wrong in thinking that he has *justification* for the existence of an external world. But it also involves realising why this needn't be a problem for our everyday practices: we are forced into *some* line of inquiry and, provided that there is no reason to think that an alternative practice might yield better results epistemically (i.e. provided there is no extant reason as to why one of our presuppositions might be false), then we may proceed legitimately in our inquiries.

References

- McGinn, M. (1989) *Sense and Certainty: A Dissolution of Scepticism*, Oxford: Blackwell
- Moore, G. (1939) ‘Proof of an External World’, *Proceedings of the British Academy* 25, 273-300
- Pritchard, D. (forthcoming) ‘Wittgenstein on Scepticism’ in the *Oxford Handbook to Wittgenstein*, (ed.) M. McGinn, OUP
- Wittgenstein, L. (1969) *On Certainty*, (eds.) G. E. M. Anscombe & G. H. von Wright, (tr.) D. Paul & G. E. M. Anscombe, Oxford: Blackwell
- Wright, C. (2003) ‘Wittgensteinian Certainties’, in *Wittgenstein and Scepticism* (ed.) D. McManus, London: Routledge

Intentionality, Intending and Moral Responsibility

Wesley H. Bronson
Princeton University

Many philosophers point to the so-called “doctrine of double effect” to show the tension between intending, intentional action and moral responsibility. The classic example goes as follows: a bomber has been given the assignment to bomb a munitions plant which is producing weapons used against his country, so he must go and bomb the plant. In bombing, however, he will undoubtedly kill innocent civilians in the nearby area of the blast. In this case, the bomber intends to destroy the plant (he desires its destruction), and he intentionally bombs it (he brings about the known consequence of his action). But while he also intentionally kills innocent civilians, he may not intend to kill them. The debate is thus whether or not the bomber is morally responsible for the death of the innocent civilians whom he intentionally kills but doesn’t intend to. With this scenario in mind, I argue in this paper that simply the fact that a consequence is brought about intentionally is insufficient for an agent to be held morally responsible, and further, that intending and being causally responsible for a consequence are often necessary to be morally responsible.

Many argue that an agent intentionally brings about any known consequence of the action he performs. So if an action will result in consequences C_1 , C_2 and C_3 , and the agent knows that a particular action will result in all three consequences, then in performing the action, the agent ostensibly intentionally brings about those consequences. There are two advantages to defining action in such a way. First, it is very simple. If an agent chooses to perform an action, then any known consequence is intentionally brought about. Second, issues over moral responsibility can plausibly be settled easily. An agent can plausibly be held morally responsible for bringing about any known consequence.¹ After all, he knew the consequence would result, and he performed an action that would bring it about. So perhaps an agent ought to be responsible for anything brought about intentionally.

Let's say I plan on burning down an apartment building. And let's say that I know that there will be various consequences of burning it down. I know many people in the building will die. I also know that the flames from the building will cause nearby structures to be somewhat damaged. And finally, I know that the fire department will spend a lot of time and money to put out the fire, probably even causing injury to several fire fighters. Given that I know ahead of time that my action (burning down the apartment building) will result in several different consequences (deaths, structural damage, etc.) it can be said that I intentionally bring about all those consequences. The simple view of intentional action mentioned above might simply say that since the consequences were intentionally brought about, I am morally responsible for all of them.

This position, however, is overly simple, and fails to take into account the mental state of the agent. Instead, the position I will use in this paper, therefore, is similar to the more complex position accepted by Michael Bratman. Rather than group all types of action for which there is a known consequence to be an intentional action, Bratman distinguishes between two types. There is a difference between *intending* a consequence and *intentionally* bringing about that consequence.² Like the previous description, any action which brings about a known consequence is performed intentionally. But what if I must bring

1 Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Ed. J Burns and H. Hart. London: Methuen, 1970, p.84.

2 Bratman, Michael. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press, 1987. Chapter 10.

about a known consequence as a side-effect of bringing about another consequence? What if I know a consequence will result, but I do not want that consequence, yet perform the action anyway? The simple view of intentional action will say that because it is a known consequence, I want to bring about that consequence. But what can be said of my mental state in which I simply do not want a certain consequence to result but know it will inevitably happen if I perform the action? In that sense, I *intend* to bring about only one of the consequences, but I nevertheless *intentionally* bring about all of them. In short, if an agent wants to perform a given action A because he wants consequence C₁, but he knows or is justified in believing that doing A will also result in consequences C₂ and C₃, then in doing A the agent can be said to intend C₁ but intentionally cause C₁, C₂ and C₃.

Many philosophers point to the so-called “doctrine of double effect” to show the tension between intending, intentional action and moral responsibility. The classic example goes as follows: a bomber has been given the assignment to bomb a munitions plant which is producing weapons used against his country, so he must go and bomb the plant. In bombing, however, he will undoubtedly kill innocent civilians in the nearby area of the blast. In this case the bomber intends to destroy the plant and he intentionally bombs it. But while he also intentionally kills innocent civilians, he may not intend to kill them. The debate is thus whether or not the bomber is morally responsible for the death of the innocent civilians whom he intentionally kills but doesn’t intend to. With this scenario in mind, I will argue in the remainder of this paper that simply the fact that a consequence is brought about intentionally is insufficient for an agent to be held morally responsible, and further, that intending and being causally responsible for a consequence are necessary to be morally responsible.

There is one exception which must be pointed out. Even those who would establish that any consequence performed intentionally results in the agent being held morally responsible would concede that an agent can be held morally responsible for actions that are not performed intentionally. That is, intentionality may be a sufficient but not a necessary condition for moral responsibility. Actions may be performed unintentionally through negligence for which we are still responsible. In cases of negligence, an agent can—but not necessarily— be held responsible for a consequence that he neither intended nor intentionally brought about. Take a drunk driver, for example. If he carelessly, in his

drunken state, runs over and kills a pedestrian, few would deny that he ought to be held morally responsible for the death of the pedestrian.³ Note, however, that the driver did not intend for the death to occur, nor did he intentionally run over the pedestrian. In the case of negligence, he merely is causally responsible, yet nevertheless remains morally responsible. That is not to say that an agent ought to be held morally responsible in all cases of neglect. If I flip a switch that looks just like a light switch, thinking it to be a light switch, and instead it turns out to be an alarm, how can I be held blameworthy for flipping the alarm? I may be causally responsible for doing so, but since I had no possible way to know that the alarm would go off, I cannot be held morally responsible for the action.⁴ So while there are indeed times when a negligent action can result in the agent being held morally responsible, it need not be the case if, for example, there was no possible way the agent could have known the consequences that he ought to be held responsible there as well.

Similarly, Carl Ginet claims that there are indeed cases when an agent both *should* have known certain information and can then be held responsible for not knowing it, even if the consequence was brought about unintentionally. If Simon went to flip a light switch, but instead of looking like a normal switch it looked identical to a fire alarm, and upon flipping the switch the fire alarm went off, it seems rational, Ginet claims, to hold him morally responsible for making the mistake.⁵ There are certain times when evidence ought to be compelling enough that the information—in this case that the switch was, in fact, a fire alarm and not a light switch—should simply become apparent. Under these situations, even if the individual remains oblivious to the situation, he can still be held responsible. As Ginet concedes, “a certain amount of indignation towards him, for his causing the alarm to go off, would be deserved (though, of course, not as much as if he had intentionally set it off).”⁶ That is, we may be less angry at the agent for his lack of intention, but he is ostensibly no less blameworthy than an evil agent. Cases of negligence, therefore, when there is a lack of practical reflection on the action and its consequences admittedly seem to undermine my thesis.

3 Mele, Alfred and Steven Sverdluk. “Intention, Intentional Action, and Moral Responsibility.” *Philosophical Studies* Vol. 82, (1996). p. 269.

4 Ginet, Carl. “The Epistemic Requirements for Moral Responsibility.” *Philosophical Perspectives 14: Action and Freedom*, 2000. Ed. James E. Tomberlin. Boston: Blackwell Publishers, 2000. p. 269.

5 Ibid., 271.

6 Ibid., 271.

My aim, therefore, must be somewhat clarified. Earlier I said that my aim was to show that establishing that an action was performed intentionally is insufficient to establish moral responsibility and that intending is a necessary requirement for moral responsibility. In light of cases of negligence, the thesis must be somewhat qualified, and I admit that maintaining intending as a necessary condition for moral responsibility only holds in cases where the agent reflects on his actions before doing it.⁷

Let us finally turn to intentional action. Why in the double-effect bomber example would someone be inclined to hold the bomber morally responsible for the death of the innocent civilians? Perhaps it is because the action was performed intentionally. Initially, the argument seems intuitively plausible. If the bomber could foresee the consequences and knew that the children would be killed, and he nevertheless bombed the munitions plant, then the fact that the action was performed intentionally would make him responsible for the deaths. But under more careful scrutiny, the fact that the action was performed intentionally is insufficient to establish moral responsibility. Let us look at a simple case of coercion to show that this is the case. Suppose Smith threatens to burn down Bob's house unless he throws a rock through a nearby window. Bob loves the building, however, and strongly opposes breaking the window. Stampeded by the threat, Bob unfortunately throws the rock, and the window breaks. Can he be held responsible for breaking the window? It seems that he intentionally breaks the window, but he is not morally responsible for doing so. Establishing that an action is performed intentionally is therefore not sufficient to establish moral responsibility.

There is a problem with the above example, however, and an obvious objection arises. Although in the case provided the individual did act intentionally and he can't be held morally responsible, he had no alternative possibilities. We tend to think that if an agent has a choice of whether or not to perform a given action, he can be held responsible for his choice. Carl Ginet comments, for example, that an individual must at some point have a choice of whether or not to perform a given action if we are to hold him morally responsible. As he defines a preliminary

7 My purpose here is to exclude cases of negligence. Negligence represents a unified and widespread class of phenomena. In a case of negligence, an agent can reflect on his action before doing it, but for some reason fails to do so. So we cannot say an agent must always intend a given consequence to be held morally responsible. My aim is to show now that in cases where the agent *does* reflect, then intending is a necessary condition for moral responsibility.

definition of the *could-have-done-otherwise* (CDO) condition: “until t_1 it was open to S not to make movement M then or any other movement that would bring about H.”⁸ That is, at a given time (t_1), an individual (S) must have a choice of whether or not to perform a certain movement (M) that will bring about a certain consequence (H) if we are to hold him responsible. In the Bob and Smith example, Bob (S) does not have the choice of whether or not to throw the rock (M). He is stampeded by the threat, so he could not have done otherwise. Those who hold that moral responsibility requires that the agent intentionally bring about that consequence will say that the CDO condition was not met in the coercion example. Since he had no alternatives, of course he can’t be held morally responsible!⁹

As an interesting note, under similar conditions, we might ask what else *would* be required for Bob to be held morally responsible. Let’s say that again, Bob is threatened by Smith, but this time, Bob hates the building and wants nothing more than to cause damage to it. He willingly throws the rock, and the window breaks. Is Bob morally responsible for breaking the window in this second case? It seems we are inclined to think yes, he is responsible. Despite the threat, Bob intended to break the window and acted intentionally. The only difference between the two cases is that in the first, Bob did not intend to break the window, but in the second case, he did intend the consequence. We might be inclined to say that intending is a necessary condition in this case for establishing moral responsibility.¹⁰

The objection over alternate possibilities is legitimate, so let us therefore turn to a case where the agent intentionally performs an action where he had alternate possibilities, yet is still not morally responsible for the consequence. Such conditions could plausibly satisfy the previous objection. Mele and Sverdlik consider a doctor who must cause a patient some degree of pain during a procedure. The doctor does everything he can to minimize the pain, and he wishes there could be no pain at all, but it is an inevitable consequence of the procedure. They claim that

8 Ibid., 268.

9 See also, Frankfurt, Harry G. “Alternative Possibilities and Moral Responsibility.” *The Journal of Philosophy*. Vol. 66, No. 23, (Dec 4, 1969), pp. 829-839.

10 As Frankfurt similarly notes, if the individual were stampeded by the threat, we would not hold him responsible for the consequences of the action. But if the threat plays no role in leading the individual to action then he performs the action of his own volition. In this case, while he had no alternate possibilities, we can nonetheless hold him morally responsible.

“insofar as an agent who is A-ing is neither aiming at A-ing nor trying to A, either as an end or as a means to (or constituent of) an end, she is not intentionally A-ing.”¹¹ Mele defends his dentist by claiming he did not specifically aim at causing the pain, and the action was thus non-intentional.¹² But what if an agent does aim at causing pain? Can his actions still be vindicated in some way?

As I will show, it is entirely possible for an agent to aim at causing pain—thus intentionally bringing about that specific consequence—yet not be held morally responsible for that consequence. At first, this statement may sound intuitively implausible, so an example is necessary. Andrew is standing with his friend Bob, both of whom have lived moral lives. Because of unfortunate circumstances, however, Bob is in a position where unless someone intervenes, he will die. Let’s say that a wild, poisonous snake is about to bite him. Andrew recognizes that the snake is about to bite Bob, and sees no way to save him except by stomping on his foot, thus causing Andrew pain which will cause him to get out of the way of the snake.¹³ Causing Bob this small pain seems to be the only way to make Bob get out of the way of the snake. In such a situation, Andrew intentionally brings about Bob’s pain by stomping on his foot. And Mele and Sverdlik would agree that the action was intentional, for Andrew aimed at causing Bob pain. Nevertheless, would we really hold Andrew morally responsible for the pain Bob experienced? After all, he saved Bob’s life by doing so.

There are now two ways in which we may analyze Andrew’s action. On the one hand, we may say despite the consequence he is still morally responsible. On the other hand, we may say he is not morally responsible for either choice he makes, including causing Bob’s pain. Let’s first look at the former situation. It is possible that Andrew is, in fact, morally

11 Ibid., 274. (In performing “intentional action,” Mele and Sverdlik imply that the individual does not intend the consequence. My aim here is to show that the middle ground of non-intentional action does not vindicate the agent in any way.

12 Two ideas must be clarified here. First, they describe intentional action in such a way that the agent does something intentionally when he intends the consequence. Second, the authors vindicate moral responsibility when the agent does not aim at causing the patient pain, or when the pain is a means to a better end. I will discuss this latter consideration later, but for now we are not concerned with consequentialist reasoning. For the example at hand, we will focus on the idea that the doctor does not aim at causing the pain.

13 Admittedly some may be inclined to say that this example seems unrealistic. It is probably true. But for the sake of the example, let us just assume that stomping on the foot is the only way that Andrew can get Bob to get away from the snake. Or, since Andrew needs to think fast and does not have time to consider other options, we can also assume he justifiably sees it as the best option.

responsible for the pain. And as we will see, his being morally responsible results in a counterintuitive conclusion. Some may argue that regardless of the positive gain of saving Bob's life, Andrew should nevertheless be held morally responsible for the pain. Given the situation, Andrew has infinite possibilities, but in terms of his approach to Bob, he merely has two. On the one hand, he can choose to stomp on Bob's foot, intentionally causing pain and thus saving Bob's life. On the other hand, he can perform any other action, all of which will result in Bob's being bitten by the snake and his subsequent death. Given Andrew's ability to either save Bob's life or permit Bob to die, how ought we to look at his choice of action? Admittedly, some may debate the difference between causing some harm and letting some harm merely occur. In this situation, Andrew could actively stomp on Bob's foot, or he could do nothing and permit him to be bitten by the snake. Someone could therefore plausibly argue that if Andrew actively caused the harm, he can be held responsible, while if he permitted Bob to get bitten by the snake, he would not be causing any harm. I do not agree with this claim, nor will I debate the issue over causing a harm to occur versus failing to prevent a harm. For the case at hand, the salient feature of the example is that a harm will occur regardless of what Andrew does. What is up to him, however, is which harm will occur. To save Bob's life, Andrew needs to perform only a small action which will, unfortunately, cause Bob harm. Because the consequence of saving Bob's life so greatly outweighs the minor inconvenience of the pain Bob will incur for that consequence to obtain, it should be obvious that Andrew ought to cause Bob the minor harm and save his life. And it would seem wrong to hold Andrew blameworthy for Bob's pain, for in causing the minor, temporary harm, he saved Bob's life. This, however, is not to say that he is not morally responsible for the pain. He may be morally responsible but nevertheless not be blameworthy, for we commend his saving Bob's life. It may even be that if Andrew ought to cause the harm, then he is positively morally responsible for doing so.

Perhaps then Andrew is doomed to be morally responsible for his choice regardless of what outcome ensues. He can be both causally responsible and morally responsible in this example. It may be asked: regardless of the choices he had to make, did Andrew not aim at stomping on Bob's foot and intentionally causing him pain? It might seem that we are forced to concede that Andrew *is* morally responsible, but he made a good choice nevertheless. If he chooses to stomp on Bob's foot, then he intentionally caused Bob pain. He should therefore be morally responsible

for the consequence. Similarly, if permitting something bad to happen is as bad as causing it to occur—and we know that Andrew can easily prevent a terrible thing from occurring—we hold Andrew responsible for Bob's death through snake bite even though, as previously mentioned, he is not blameworthy. In our case then, Andrew might be held morally responsible for either consequence that occurs. What is striking here is that we tend to think that if an agent had alternate possibilities and is morally responsible, then he is either praiseworthy or blameworthy. And we have said that Andrew should be commended for causing Bob the minor harm, not criticized for his action. So even though he is morally responsible, he is nevertheless praised for his act. If Andrew is morally responsible, his being praiseworthy is itself a counterintuitive conclusion of the example at hand, possibly negating an idea that we generally take for granted—that if I'm morally responsible for causing pain then I'm not praiseworthy for causing pain.

But we are not committed to the above claim, however, for Andrew is not necessarily morally responsible for causing Bob the minor harm. It may be argued that Andrew is not morally responsible for either action he performs. If we assume that Andrew is a morally upstanding individual and has lived a moral life, yet somehow circumstances put him in this situation, and we recognize that he did not knowingly put himself in this situation, Andrew is an innocent individual. Rather than hold him responsible for either consequence, he ought to be innocent regardless of which choice he makes, for it is merely the circumstances that force him to choose between one of two bad outcomes. Andrew recognizes that it is unfortunate he must make this choice between causing Bob a harm at the least, or observing his death, so he intends for Bob to not experience any pain at all. This, however, is not an option. Given the situation, therefore, we may commend Andrew for saving Bob's life in causing the pain, but we cannot hold him morally responsible for causing Bob the minor harm, even if it was performed intentionally.

If we are right that Andrew intentionally causes Bob's pain and is causally responsible, but is not morally responsible, what other condition would have to hold for Andrew to be morally responsible for the consequence? What other feature of Andrew's mental state would we have to add so that we can unquestionably hold him responsible? It should be apparent that only if Andrew desired for Bob to experience the pain would he be responsible for stomping on his foot. In the situation above, we assume that Andrew does not wish for Bob to

experience any pain at all, but since Bob's only chance at survival forces Andrew to intentionally cause him pain, Andrew cannot avoid a harmful consequence. But if Andrew gets some enjoyment out of Bob's pain, then the situation is very different. Andrew's intending to cause Bob pain in this situation is a necessary condition for moral responsibility. What previously vindicated Andrew's action in the seeming dilemma was that although he intentionally brought about a consequence, he did not want to cause the pain in the first place. So if we remove that component of Andrew's mental state and turn him instead into an agent who desires Bob to experience the pain of having his foot stomped on, then we are inclined to hold him morally responsible for Bob's minor harm. Indeed, what vindicates moral responsibility is the lack of intending a specific consequence even if the action is performed intentionally.

Mele and Sverdlik raise an issue that makes us question when we consider an action to have been performed intentionally, and thus force us to look at another type of case for moral responsibility. Referring to a problem first raised by Ronald Butler, they quote the following problem:

If Brown in an ordinary game of dice hopes to throw a six and does so, we do not say that he threw the six intentionally. On the other hand if Brown puts one cartridge into a six chambered revolver, spins the chamber as he aims it at Smith and pulls the trigger hoping to kill Smith, we would say if he succeeded that he had killed Smith intentionally. How can this be so since the probability of the desired result is the same?¹⁴

Butler's problem is not entirely complicated. Brown has a one in six probability that the dice will land six up, and a one in six probability that the gun will shoot a bullet and kill Smith. To make the problems closer to one another, Mele and Sverdlik create their own situations similar to Butler's case. In the first, Brown wants to kill Smith (K-ing). In order for this to happen, however, Brown must throw a dice so that it lands six face-up, causing a chain of events which he knows will result in an explosion that kills Smith. So in throwing the dice, Brown has a one in six probability that he will accomplish his goal and Smith will die. In the second case—which is supposed to more closely resemble the non-moralistic case of simply getting a six in a game of dice—Brown wants to gain membership to a given fraternity (G-ing). Like in the first case,

¹⁴ Ibid., 279.

Brown can only gain membership by rolling a six, so the chance that he gains membership is one in six. The control Brown has in each case is the same, so is it possible that one action was performed intentionally and the other was not?

Mele and Sverdlik's cases are accurate parallels of Butler's original problem. So Butler might be inclined to say that Brown did not intentionally G but he did intentionally K. Why might he be inclined to think this? Since there is only a limited control Brown has over the possibility of throwing a six, it is difficult to say that he did it intentionally. Yet Butler claims we are inclined to say that Brown killed Smith intentionally. In light of Mele and Sverdlik's examples, however, we see that there is no reason to be committed to this claim. That is, there is no reason to think that one action is intentional and the other is not. We must be either committed to the fact that Brown neither K'd nor G'd intentionally, or that he did intentionally perform both actions. As they point out, there is nothing in either of the cases to make it seem that it is plausible that one was more likely to be performed intentionally than the other. Brown, in each case, merely threw the dice hoping for a certain outcome, but while having equal control over it. In the end, however, they claim that Brown ought to be held responsible for the death of Smith, but that the moral responsibility does not lie in the fact that he performs the action intentionally. In doing so, however, they miss the most salient argument for why we hold Brown responsible. So what do Mele and Sverdlik overlook?

The answer lies in a deeper analysis of intentionality and intending. What they ignore here is *why* we are still inclined to hold Brown responsible, aside from the fact that he is partly causally responsible. As I argued earlier, I maintain that intending is a necessary condition for moral responsibility. What seems most important in the Brown case is that Brown *intended* for the death of Smith to occur. And although he had only minimal control over the outcome, he was as causally responsible for the event as he could be. It is the fact that he intended the outcome that causes him to be responsible for the death. Similarly therefore, Brown is responsible for throwing a six, whether or not he intentionally did it. In killing Smith, Brown intended to throw a six. We again support the thesis, therefore, that under certain situations, intending a consequence is a necessary condition for moral responsibility.

Someone can respond to all of this, of course, by claiming that even if Brown, for example, did not intend to bring about Smith's death,

his rolling the die intending for Brown to survive makes him no less blameworthy for Smith's death. Perhaps then Brown is responsible for the action that he does not intend, and I am wrong to claim intending is necessary for moral responsibility. On the one hand, this may seem like a plausible defense. It seems like we have created a situation in which Brown is morally responsible despite not intending for the consequence to obtain. But we must ask a question about an agent who performs an action knowing a certain consequence may occur but performs the action hoping for it not to occur. Namely, why did he perform the action in the first place? If it turns out that Brown had no reason at all to kill Smith and he did not want Smith to die, then there was no reason to roll the dice knowing that Smith might die. If this is the case, then of course we should hold Brown responsible. Performing some action that we know will lead to a consequence simply on a whim—even if we don't intend that consequence—cannot negate moral responsibility. On the other hand, if Brown were to throw the dice intending to throw a five, but had some other motive for throwing the dice, for example the possibility of some greater good, then we would not hold Brown responsible. As mentioned earlier, even if an agent is morally responsible, he need not be blameworthy for the consequence. Andrew might be morally responsible for the pain Bob experiences in having his foot stomped on, but Andrew is not blameworthy for causing the pain. So it is still plausible that even if Brown intends to throw a five, he can still avoid being blameworthy, even if he is causally responsible and intentionally brings about Smith's death.

At this point, I hope it has been shown that under certain circumstances, being causally responsible and intentionally bringing about a consequence is insufficient for holding the agent morally responsible for that consequence. In these situations, what enables us to hold the agent responsible is only when he intends that particular consequence. Let us now look back at the double effect bomber example. Under more careful scrutiny, it is no different than the doctor example. In each situation, the agent is causally responsible for the consequences of his actions. And both the doctor and the bomber intentionally cause some harm. The doctor intentionally causes his patient's pain, and the bomber intentionally kills the innocent civilians. But if I have shown thus far that doing something intentionally and being causally responsible for the consequences are insufficient conditions for moral responsibility, then the bomber is in the same position as was the doctor. And if we claim that the doctor is not morally responsible for the pain his patient incurs,

then similarly the bomber cannot be held responsible for the death of the innocent civilians. He is causally responsible and intentionally kills the innocent civilians, but he does not intend to cause their death, and as we have shown, when these conditions are satisfied, the agent is not morally responsible. Only when the bomber intends to cause the death of the innocent civilians and then bombs the munitions plant can he be held morally responsible for his actions.

The extent to which this idea can be pushed further, however, is unclear. Let's say that a new poison was found, but it was unknown how long it takes to kill an individual. So for the sake of knowing this information, I administer the poison to someone. I do not want the person to die, but know it is an inevitable outcome of administering the poison. And the only way to know how long the poison takes to kill the man is to perform this experiment. In this situation, I am causally responsible for the death of the man, I intentionally kill him, but I nevertheless do not intend for him to die. If there were a way to find out how long the poison takes to work without killing him I would perform that experiment instead, but unfortunately, there is no other option. While I wish the man could survive, if I am to find out how long the poison takes to work, I must intentionally kill him. At first, it seems that I must be morally responsible for the death of the man. After all, I poisoned him simply to find out how long the poison would take to work. There is no redeeming feature here (for example, that his death will allow me to create an antidote that will save hundreds). Simply put, I kill him to find out some useless information. Can I really be morally innocent here? Admittedly it feels wrong, but if what I have said previously holds true, I am forced to accept that I am not morally responsible for the man's death. While someone may argue that we are inclined to accept the bomber as not morally responsible for the death of the innocents because destroying the munitions plant seems to be an acceptable cost, such a defense would miss the purpose of this paper. Appealing to consequentialist reasoning would precisely avoid the jarring nature of the counterintuitive conclusions about the relationship between intentionality, intending and being morally responsible. The fact that the destruction of the munitions factory is a worthwhile target does not vindicate the bomber in killing the innocents. In fact, a consequentialist could care less about this entire paper's discussion on intention. So appealing to such reasoning defeats the purpose of differentiating doing something intentionally and intending a consequence. We vindicate the bomber on the basis that he

did not intend to cause the death of the innocent civilians. The fact that some greater good could be achieved through his action is not germane to the debate. It is simply because although he intentionally kills them, he does not intend to kill them. So while I intentionally kill the man to find out how long the poison takes to work, it is entirely plausible that if I see his death as an unfortunate consequence of the poison experiment and do not intend to bring about his death but know it will occur, I am not morally responsible.

In the end, those who claim that an agent who brings about a known consequence ought to be morally responsible for that action ignore a major feature of the mental state of the agent. For there are times when we do not want to perform an action or do not want to bring about a given consequence, but know it must be done. Under these situations, while we intentionally bring about the known consequence, it seems we can be innocent. But while we may have found a plausible argument for allowing an agent to perform intuitively horrible acts and nevertheless be innocent, it does not seem that we must allow the agent to remain completely blameless. After all, if I kill a man for the sake of finding out how long a poison will take to work, regardless of whether you hold me morally responsible or say there is a redeeming feature—that I did not intend the consequence—I did nevertheless kill the individual. Some may be inclined to say that the position is implausible. We have a universal intuition that the individual who kills a man as an unintended consequence of some trivial act he did intend is definitely morally responsible. There are two responses to such an objection. First, the intuition about moral responsibility lacks a certain normative force. It is not so clear why these intuitive gut responses to the conclusions of the paper provide a justifiable objection. Second, the intuitions themselves seem linked to consequentialist reasoning. The detractor of the view presented here will respond by noting how the cost of the death cannot possibly outweigh the minor gain of the experiment. If we are to support a degree of consequentialist reasoning in our deliberations, then the objection would be strong. But if we do not include such reasoning, then it is the burden of the detractor to explain what it is specifically that makes the view implausible.

References

- Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Ed. J Burns and H. Hart. London: Methuen, 1970.
- Bratman, Michael. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press, 1987.
- Frankfurt, Harry G. "Alternative Possibilities and Moral Responsibility." *The Journal of Philosophy*. Vol. 66, No. 23, (Dec 4, 1969), pp. 829-839
- Ginet, Carl. "The Epistemic Requirements for Moral Responsibility." *Philosophical Perspectives 14: Action and Freedom*, 2000. Ed. James E. Tomberlin. Boston: Blackwell Publishers, 2000. 267-277.
- Mele, Alfred and Steven Sverdlik. "Intention, Intentional Action, and Moral Responsibility." *Philosophical Studies* Vol. 82, (1996) pp. 265-287.

The *Natural-Artificial* Distinction: A Limit Concept for Truth and Reality and its Application Across Matrices

Avi M. Miller
Princeton University

The aim of this paper is to attempt to reconcile two seemingly conflicting intuitions regarding the nature of reality. The first is that the experience of 'a brain in a vat' cannot represent reality. The second is that every person experiences the world differently; no two people share the same reality or conception of the world. The former implies that there is an objective reality while the latter puts that belief into question. The brain-in-a-vat intuition is further challenged when we consider its theological analogue. Would our brain-in-a-vat experience still be unreal if God were the creator of the matrix rather than an evil genius or complex machine? Many religious and theological perspectives cohere with such a vat model and it would be strange to think that holding such religious views implicates a disbelief in the reality of the world. In hopes of answering these questions, I will present an argument for the unreality of vat experience which, unlike other arguments previously offered, respects a broad and liberal notion of reality, one that allows for the possibility of intersubjectivity within human experience. My purpose is not to define reality or prove the reality of any particular experience or matrix, but rather to define a limit concept of reality and prove strictly that vat experiences cannot be real. I believe that such a limit concept can be captured by introducing the *natural-artificial* distinction, although not without its own set of complications and worries.

The aim of this paper is to attempt to reconcile two seemingly conflicting intuitions regarding the nature of reality. The first is that the experience of ‘a brain in a vat’ cannot represent reality. The second is that every person experiences the world differently; no two people share the same reality or conception of the world. The former implies that there is an objective reality whilst the latter puts that belief into question. The brain-in-a-vat intuition is further challenged when we consider its theological analogue. Would our ‘brain in a vat’ experience still be unreal if God were the creator of the matrix rather than an evil genius or complex machine? Many religious and theological perspectives cohere with such a vat model and it would be strange to think that holding such religious views implicates one to disbelieve in the reality of the world.

In this essay, I will begin by briefly contextualizing the brain-in-a-vat hypothesis. I will distinguish what I shall call basic matrices (“worlds”) from vat matrices and show how from such a distinction we can present a more formal argument for the unreality of vat experience. I will then critique the argument by appealing to the plausibility of intersubjectivity of human experience. In light of such a critique, I shall present the strongest case I can make *against* the unreality of vat experience, that is, for the reality of vat experience despite our intuitions and arguments otherwise. I hope to successfully challenge that argument and present a new one for the unreality of vat experience which, unlike the first, respects a broad and liberal notion of reality and that allows for the possibility of intersubjectivity within human experience. In such a manner, I hope to reconcile our two competing intuitions and, in so doing, also our theological worry. My purpose, to be clear, is not to define reality or prove the reality of any particular experience or matrix, but rather to define a limit concept of reality and prove strictly that vat experiences cannot be real. I believe that such a limit concept can be captured by introducing the *natural-artificial* distinction, although not without its own set of complications and worries.

Let us begin, then, with the premise accepted by most philosophers (with the notable exception of direct realists like Thomas Reid) that human experience is mediated by our senses. If this is the case, there arises one of the oldest and most contentious questions in philosophy: What is the *source* of our sense data? What lies on the other side of our senses? Descartes argued for the existence of an external material world. Berkeley, in contrast, argued against an external material world and

instead argued that God placed our sense data in our minds. One famous skeptical response is the possibility that we are merely brains in a vat. Our sensations and experiences are just brain stimulations administered by machines or computers while, in truth, we are just sitting in a vat.

The brain-in-a-vat scenario can be easily misunderstood. For unlike the theories of Descartes or Berkeley, the brain in a vat 'theory,' if we can call it that, presupposes a human world outside the vat. The classical 'vat' story goes something like this: man lives in the world and then one day gets plugged into a computer simulation. For the purposes of this essay, it does not matter whether or not a person ever actually experienced the world outside the simulation or if it is even within his power to do so. The fact remains that in the case of the brain in a vat there is an original, *human* world denied to *humans*. This is not the case with either Descartes or Berkeley. Although *God* may, for either thinker, exist in a 'world' outside of human sensation, such a world cannot conceivably be one occupied by *humans*. Thus we must be careful to note that the brain-in-a-vat hypothesis analyzes the source and status of only the simulated sensations, not the sensations of a person had he lived in the actual world presupposed. The brain-in-a-vat hypothesis does not offer an explanation of the original source of human sense data (i.e. *before* they were plugged into the machines) but rather offers a line of skeptical reasoning that puts into question whether the sense data we perceive accurately reflects the world we live in.

Notwithstanding this important distinction, the classical brain-in-a-vat theory still describes a possible world and explains the source of at least our immediate sensations if not our original ones. For the purposes of this essay, any possible network of inputs and outputs that accounts for human sensation and which ultimately produces the 'world' which we can immediately perceive through our senses I shall call a matrix. Matrices can either include or exclude a material external world. As we said above, Descartes argued that God created an external world which, in turn, is the source of our sense data. Berkeley argued that God directly places those sensations in our minds with no corresponding material external world. These are two alternative matrices, two possible worlds with different accounts of what the world is and how we perceive it. Yet just as matrices need not have an external world, they also need not have God as its creator. In other matrices, the Big Bang may explain the existence of an external world without a creator at all and evolutionary theory may explain how man adapted senses to perceive it. Alternatively,

if there was no Big Bang, another evolutionary theory may argue that humans evolved senses that originate in their biochemistry rather than in any correspondence with an external world. In the latter matrix, the world we perceive does not correspond to anything outside of us, but rather results from our neurochemistry—a Berkeleian matrix, so to speak, without a God. The table below summarizes the four basic matrices I just described.

TABLE OF BASIC MATRICES

External World	No External World
Creator of External World - God (Descartes)	God directly implants sensations into man's mind (Berkeley)
No creator - Evolution/Big Bang (Brian Ellis)	Human evolution/ neurochemistry alone accounts for sensation

Brain-in-a-vat matrices can be said to be a special, complicated form of any one of these basic matrices. Some being – whether an evil or beneficent genius, a computer, machine, or man himself – manipulated our senses (however they were naturally or divinely construed) so that they are now entirely dependent on the program. The easiest example to use is an external world matrix. Regardless of who created the external world or if there even be a creator, some being reconfigured our senses so that they no longer correspond to that external world. The same idea is true of matrices that do not involve external worlds. Some being reconfigured our senses so that they no longer correspond to their original source—perhaps God, perhaps their neurochemistry. This distinction between basic matrices and the special case of brains in a vat is important to keep in mind as this paper progresses.

Thus far all we have said concerned possible sources of sense experience. But now we encounter a new claim that, if it turns out that the immediate *source* of our sense experiences is some brain-in-a-vat computer simulation, such experiences are not *real*. The intuition for such a claim nicely follows from what was said above. Brain-in-a-vat matrices presuppose a natural, human matrix outside the vat. This predisposes one to say that the original natural matrix world represents

reality, whereas the vat-world represents some fictional, imaginary, or deceptive non-reality. The vat-world is ontologically distinct from the presupposed *real* world, and thus *unreal*, no matter how closely it may correspond to it.

To clarify the point above, compare vat matrices with Berkeley's idealist matrix. In some ways, Berkeley's matrix shares quite a lot in common with vat matrices. Both matrices have a creator; machines, computers, etc. or God. In both matrices, the material facts which correspond to human experience are not the facts we believe them to be. Inside the matrix, we believe that it is our eating a steak that causes us to taste its fine grilled flavor. In truth, however, we were sitting in a vat doing no such thing and either God or some machine stimulated our neurons in order to make us believe that we had a steak on our plate, we were eating it, and tasting it.

The crucial distinction however between the Berkeleyan matrix and the vat matrix is that the former does not presuppose some more natural human state. In Berkeley's matrix, God created man and placed him within this world *from the beginning of his creation*, and thus there was never a time 'before he was plugged in' to the divine matrix. Moreover, man in his present form cannot exist outside this world; he was born within it and cannot escape it. The reason why this distinction is so significant is that it denies the possibility of having the same intuition for Berkeley's matrix as we had earlier regarding vat matrices. For there is no pre-vat, original real world for the Berkeleyan matrix not to correspond to, that would lead us to conclude those experiences are unreal.

This is certainly not to say that from this we can conclude that human experience for Berkeley (and Descartes) attains the status of the *real*. On the contrary, all I mean to argue is that their respective matrices cannot be deemed *unreal* for the same reason as brains in a vat. In summary, brain-in-a-vat matrices are inherently different from the other basic matrices in that there is an original human world that has since ceased being the source of human sensation. This difference seems to motivate our intuition that vat experiences are not real. Whether we can affirm in the positive that our sensations in basic matrices are real is still up for further scrutiny.

Let us now move beyond our intuitions and clearly articulate and examine the standard argument for the unreality of vat experiences (or vat-worlds). The argument may go as follows: when we are plugged into a vat world, we sever the connection our senses had to the *objective* real

world that we, as a species, had always lived in. Inside the vat-world, we may go to school, have friends, get married, vacation at a tropical beach, but in reality, all we are doing is sitting in a vat. This kind of ‘objectivity of facts’ argument is the one James Pryor mentions in his article ‘What is so bad about living in the Matrix?’ No vat experience can be said to be real, because reality is not defined by the neuron messages in our brain, but rather by the objective facts outside it. We are sitting in vat, no matter what our brain thinks.

To further understand what is going on in this argument, I think it will be helpful to reconstruct it using Berkeley’s crucial distinction between human understanding and human will. Let me just warn the reader in advance that I do not intend to reconstruct Berkeley as much as merely use his conceptual tools. I believe that, in so doing, our analysis and eventual critique of the argument will be much richer. Berkeley writes, “A spirit is one simple, undivided, active being: as it perceives ideas it is called the *understanding*, and as it produces or otherwise operates about them it is called the *will*” (On the Principles of Human Understanding, §26). “Will” narrowly defined is the imagination. One could be sitting in a lecture hall and imagine himself sunbathing on a tropical beach. One can imagine snow in august or a moon made out of blue cheese. Berkeley writes,

I find I can excite ideas in my mind at pleasure, and vary and shift the scene as oft as I think fit. It is no more than willing, and straightway this or that idea arises in my fancy; and by the same power it is obliterated and makes way for another. This making and unmaking of ideas doth very properly denominate the mind active. This much is certain and is grounded on experience. (*Principles*, §28)

“Understanding” refers to perceiving the sense data that presents itself to us, and not that which we actively conjure up. We close our eyes, open them, and if we are sitting in a lecture hall those certain sensations of the room impose themselves on us: the chairs, the desks, the podium, the chalkboard. No one wills these sensations, they just appear to us. In essence, the understanding/will distinction can be said to be the sense/imagination distinction. In Berkeley’s words,

Whatever power I may have over my own thoughts, I find the ideas actually perceived by sense have not a like dependence on my will. When in broad daylight I open my eyes, it is not in my power to

choose whether I shall see or not, or to determine what particular objects shall present themselves to my view; and so likewise as to the hearing and other senses, the ideas imprinted on them are not creatures of my will. There is therefore some other will or spirit that produces them. (*Principles*, §29)

Using this distinction, Berkeley claims that our understanding is more real than the will and therefore our sense data is more real than our imagination. He argues this by stating that our understanding is much more orderly than our will; that is to say, our passive sensations much more structured than our active imaginations. Though I believe this type of ‘argument by order or design’ fails in the end, I will not defend myself here since it is only tangentially relevant. More importantly, though, we can apply Berkeley’s terminology to what we have said above and say that the understanding is more real because it represents *objective facts*, whereas the will does not. In such a spirit of reconstructing the ‘objectivity of facts’ argument, I believe I can extend this to mean that any vat-world is less real or unreal compared to the presupposed natural world. The vat-world is, by definition, a manifestation of will, whereas the presupposed world outside the vat is, in a way, the world of understanding or the world of objective facts. Some may say that the vat-world is not the product of our own will but rather the will of some other being like a machine. Still it cannot be regarded as real as the objective facts of the presupposed natural world. For if we say that our own will or imagination is less real than our understanding, certainly someone else’s imagination imposed upon us is less real than our own understanding.

I have thus presented ‘the objectivity of facts’ argument for the unreality of vat experiences. The argument, however, is vulnerable to a serious challenge. By accepting its premises, we are forced to dismiss all too quickly the possibility of intersubjectivity of human experience. What I would like to do now is not so much prove definitively such a possibility as much as make it seem very plausible. By doing so, I will have shown the need for a new argument that will prove the unreality of vat-worlds while still respecting human intersubjectivity, or alternatively a strong reason to give up the claim that vat-worlds are unreal altogether.

The crux of my objection to the ‘objectivity of facts’ argument is the fine line drawn between understanding and will. It seems that upon closer scrutiny, human beings cannot strictly separate understanding and will. Perhaps, in the narrowest of senses we can separate our wild

imaginings from the unwilling sensations of the understanding. We can distinguish our imagined fantasy of sunbathing on a beach from the brute sense of sitting in a boring lecture. But in a much broader (and I believe correct) understanding of will, stands all our beliefs, opinions, upbringing, schooling, and past experiences—all of which shape how we see and experience the world. If we can appreciate the ease with which we can distinguish the content of our imagination as we daydream in class from what is actually going on in the lecture hall, we should also appreciate the difficulty of the grayer, more common and complex cases when it is not so easy to tell the difference between how the world *really is* and how the world *seems to us* at a given moment. To a certain degree, we all see the world the way we want to see the world. Where one person sees the hand of God, another person sees chance. Moreover, we are all very selective in what we consciously perceive. As individuals, we cannot help but experience the world uniquely. I see the world, experience the world, and interpret the world quite differently than my Rwandan classmate or my best friend since childhood, or anyone else for that matter. Even if we can distinguish our imagination from that which is not imagined, it seems an almost impossible task to distinguish the sensations ‘given to us’ and our will that selects, processes and interprets them. I am not asserting here that objective brute facts are a myth, although perhaps they are. I am merely asserting that we cannot know them (or most of them), if they do exist.

To put this idea in Berkeleyan terms, in every act of the ‘understanding,’ the ‘will’ asserts itself. We can never purely ‘understand,’ that is to say we are never passively accepting sensations that are given to us without some activity of our own, whether it is selecting which sense data to perceive, interpreting that data uniquely, or some other operation during the complex process of perception. We may be able to purely ‘will’ or imagine something out of the thin blue that has absolutely no correspondence with our sensory inputs. However it is difficult to claim that we can perceive something purely without the interference of any active human processing (which inevitably alters the content). Again, the claim cannot be defended here in full, but I hope to have at least demonstrated its viability.

There is an interesting discussion found in Van Fraassen’s *The Scientific Image* that questions at least some of what has been just asserted. Van Fraassen argues that there is an important distinction between *observing* and *observing that*. He claims that *observing* means

pure physical perception whereas *observing that* requires a conceptual awareness of what that perception contextually signifies. He uses the example of a Stone Age person observing a tennis match. Certainly, they perceive the same physical things we do, but perceiving *that* they signify a tennis match would require much more information than was available to the Stone Age person. Thus we can all *observe* objective facts, even if our comprehension or interpretations of such facts are different. In Berkeleyan terms, Van Fraassen seems to be implying that indeed we can distinguish will from understanding, even if both are simultaneously present. We *understand* (or *observe*) objective facts and *will* (or *observe that*) its interpretation. As such, we can clearly identify and separate out the objective facts, i.e. the physical perception of the tennis match, from the confines of our will, i.e. the comprehension of the tennis match.

My reply to this is that differences in *observing that* may actually cause differences in simple *observing* in a way that will make it quite difficult, if not impossible, to identify the objective facts within our perceptions. Certain psychological theories of perception, for example, argue that theoretical, cognitive constructs often help guide and determine what we actually decide to selectively perceive. One study showed that linguistic differences between German and English actually create differences in ordinary perception. English has a progressive that is not available in German. In English, we can say 'he is crossing the street,' but in German we must say 'he crosses the street.' It has been shown that when a native English speaker perceives a person walking from A to B, their eyes tend to track the person's movement. When a native German speaker perceives that very same person, his eyes fixate on A and then B, without following his movement. This study has shown that linguistic differences (*observing that*) actually cause differences in observing.

It seems quite likely then that, by extension, our beliefs, experiences, assumptions, and cultural and religious heritage may also help determine simple *observing*. Psychological research supports this hypothesis as well. It seems, therefore, not so unreasonable to say that the Stone Age person might actually physically perceive a very different tennis game than the 21st century person sitting beside him. Perhaps the Stone Age person is concentrating on the fans clapping in the second row rather than the players and the ball. Or perhaps he fails to see the competitive or athletic aspects of the game or the stakes involved or the rules that govern the game. In summary, it still seems that we cannot abstract away from our embedded selves and distinguish *observing* from *observing that*.

What I have tried to do is illustrate cases where it isn't clear that there is an objective fact of the matter, or that if there is, subjective experience makes it impossible to determine what that might be. I would like to suggest that human experience really is intersubjective more generally. Some may find the treatment of the matter way too hasty and they may argue that the examples I bring are not illustrative at all. Perhaps they are right. In my defense, I cannot possibly give an account here that generalizes this idea sufficiently to imply that most of our experiences fail to represent objective fact. That requires too much work that goes well beyond the confines of this essay. In the end, I leave it up to the reader to decide whether he believes human experience is intersubjective. For those who are convinced it is not, then perhaps the work for you is done. Yet for those who believe that a definition of reality should at least leave conceptual space open for the possibility of incorporating elements of human intersubjectivity, I ask that we press on. I only remind the reader that even if an account for human intersubjectivity is now elusive, it is far from impossible.

To complete this discussion and bring the point home, let us return now to the 'objectivity of facts' argument. The argument claims that vat experiences are not real because they do not represent the objective facts outside the vat. But if we accept the possibility of intersubjectivity within human experience, then even our experiences outside the vat do not represent the objective facts, if objective facts even exist at all. This is so because, as we said above, our experiences even outside the vat are inescapably the products of both our understanding and will. How then can we explain why vat experiences are any less real than experiences outside the vat?

We are thus left with three options. One, we can claim that despite everything we have said to the contrary, man experiences and perceives objective facts. In that case, then there would be no problem in continuing to use the objectivity of facts argument to prove the unreality of brain-in-a-vat matrices. Two, we can continue to entertain the possibility that human experience is intersubjective and then, in order to remain consistent, claim that vat experiences indeed also represent reality. Three, we can continue to entertain the possibility that human experience is intersubjective and then search for a new argument to show that vat experiences cannot represent reality. The first option will not be further considered in this paper. I will now turn to consider the second option in more detail before dedicating the remainder of this essay to the

third option.

The second option embraces the idea that vat experiences are real just as our subjective experiences outside the vat are real. This essay thus far seems to have taken the majority opinion for granted that vat experiences should not be considered real. But this might be too quick, for perhaps what is important or real is merely the sum of my experiences alone. Pryor partially makes this point (although does not subscribe to it) by using the example of eating steak. Why should it matter at all that a brain in a vat is not actually eating a steak in his vat, but merely having the experience of eating one due to the stimulation of certain neurons in his brain? Why can't the simulated experience of eating a steak be just as real as when there is an actual steak? More generally, why should reality be contingent upon the source of the experience? This is an absolutely crucial question that I will return to. But what I am suggesting now with the steak example is that, from this perspective, vat experiences may turn out to be real, contrary to what we have said all along.

To make the strongest case I can for the reality of vat experiences, I wish to consider Brian Ellis' remarks in his essay *What Science Aims To Do*. Although Ellis does not directly address brains in vats, what he does say about the nature of reality and truth will ultimately become very relevant to our purposes. In short, Ellis claims that reality and truth are dependent on the types of beings we are; they can only be understood within the framework of our limited human epistemic abilities. For us, reality is human reality and truth is human truth. What I hope to suggest is that if this is so, then perhaps brains in vats live in their own vat reality, different from the reality of machines or whoever is operating the program from outside, just as humans experience reality differently from bats. Likewise, perhaps brains in vats can attain vat truths different from the machine truths outside the vat, just as human truth may be very distinct from some alien truth. If Ellis is right that "it is enough if there is a way that the world is *for us*," then there may be a strong case for the reality and truth of vat experiences (p. 73).

Let me now attempt to develop this line of thinking in greater detail. Ellis presents an account of the subjectivity of human experience very different from the one I offered above. Ellis believes, like most realists, that reality is not just some construction of ideas and sense impressions but exists independently of anyone's knowledge or beliefs about it. "One cannot change the basic structure of the world by changing one's beliefs about it" (p. 70). For him there exists an external material world that is

the source of our sense experiences. So far it seems fair to conclude that such claims endorse the view that objective facts exist.

However in this case, unlike most (metaphysical) realists, Ellis further argues that “the way the world is is relative to the sorts of beings we are” (Ellis, p. 71). What he means is that *our understanding and experience* of reality is necessarily dependent on the sorts of beings we are – our biology, evolutionary history, epistemic values, etc. This fact does not make it wrong to believe what we do about the world just as it would not make it wrong to believe something different if we were different beings. Ellis writes,

I am not denying the existence of a [common] world... What I am denying is that there is any *way* that the world is independently of how it is for various kinds of beings. Different beings may have different perspectives on the world, but there is no reason to think that there is any perspective which can claim priority... [Thus] no objective stance which would define the way the world is independently of our epistemic values seems to be possible. (p. 72)

This position does not deny the existence of objective facts, only the possibility that we can ever know them. As an illustration, Ellis writes that if we ask “‘Why do things behave as if the theory T were true?’ the perfectly adequate answer is that it is because the world is, for us, a T world” (p. 73). In essence, Ellis makes a similar point I had made above concerning the subjectivity of human experience; the difference is that I argued more ambitiously for *interpersonal* subjectivity of human experience, while he is arguing for *interspecies* subjectivity of experience.

Ellis’ concept of reality, then, seems to be integrally tied to the resources available to a particular species – he allows something to be real for one species and not for another. He maintains his realist objective thesis by claiming that while reality is dependent on our experience of it, it is not causally dependent. To illustrate this point, he gives the example of moving the position of our little finger. By moving our finger, we do not *causally* change Sirius in any way, but we do change its relationship to our little finger. In the same way, if our epistemic values change, our beliefs about the world change; therefore, we also change the way the world is *for us*. But we do not causally change the nature of reality in the sense that reality would change *for other beings*. In this manner, he thus synthesizes his realism and subjectivism (or, as he calls it, relativism).

To return now to our original question concerning vat experiences,

it seems there is a compelling reason to argue that vat experiences are real. If reality is integrally tied to the resources available to a particular species, then brains in vats could be said to exist in their own reality according to the resources available to them. Moreover, as we said above, there is no absolute reality, only various realities according to the various limits of the species. If this is so, and the human species was brains in vats, then our vat worlds should represent a reality *for us*.

Moreover, we may say something analogous about truth. Thus far, we have not mentioned truth in much detail. It will suffice to say that, as a general rule, the same arguments that have been presented above concerning the reality and unreality of experience can likewise be applied to the truth and untruth of facts in that world. For example, just as above we originally held the basic intuition that vat experiences are not real, here too the immediate intuition is that we cannot attain truth inside the vat unless it corresponds to the truth outside the vat. For example, if the machines programmed the same laws of physics into the computer simulation as were found outside the vat, then there would exist truth inside the vat. But it is true solely on account of it being true outside the vat. Yet if a vat 'truth' contradicts a truth outside the vat, the immediate intuition is that the vat 'truth' cannot be considered true.

However, if we take Ellis' pragmatic theory of truth seriously, it seems a strong case can be made for the possibility of vat truth that may or may not cohere with other understandings of truth outside the vat. Ellis introduces his pragmatism by contrasting it to the correspondence theory of truth which examines truth as a "relationship between what is true and what it is true of" (p. 68). Truth, in the correspondence sense, is some metaphysical entity, independent of rational evaluation and our epistemic values. Ellis rejects the correspondence theory by the following skeptical challenge: if truth corresponds to some unknowable metaphysical truth, how do we know if we have ever attained truth or will ever attain truth, or more fundamentally, whether these metaphysical entities even exist?

Instead, Ellis supports the pragmatic theory of truth which posits no such metaphysical entities. Truth, for pragmatic theorists, is that which is ultimately justifiable, the culmination of human science guided by its epistemic values and limits. We cannot escape our own limits, nor can we ever observe the world from a nonhuman perspective. Thus we cannot hold ourselves to some standard other than human. "Truth is just the culmination of the process of investigating and reasoning about nature

in accordance with these [i.e. our human epistemic] values” (Ellis, p. 69). “Human truth is all we can aspire to,” says Ellis, while metaphysical truth seems surprisingly unattainable and incomprehensible. “If there is another kind of truth for beings of another kind, then that’s their problem. There is not and cannot be any absolute truth, and therefore there cannot be any way that the world is independently of how we, or some other kind of creature, would evaluate its beliefs about it” (p. 72).

Does the pragmatic theory imply then that truth can be attained within a vat world, and not just accidentally so? In a correspondence theory, it seems that, on the contrary, there cannot be vat truth unless it *corresponds* to some metaphysical absolute truth outside the vat. Yet in the pragmatic theory, it would seem that if all we meant by truth was that which is ultimately justifiable to us, then vat worlds can contain their own vat truths distinct from truths outside the vat. The final science of brains in a vat should be deemed truth, regardless of any *correspondence* outside the vat.

To summarize what has been said thus far, I first presented the ‘objectivity of facts’ argument for the unreality of vat experiences. After raising the objection that this argument fails to account for the possibility of human intersubjectivity, I have just put forward the strongest case I can make for the reality of vat experiences, as well as the possibility of truth within a vat world. I will now argue that this cannot be so. The third option I suggested above was to find a new argument that demonstrates why vat experiences are not real and the possibility of truth is only accidental, while still maintaining the intersubjectivity thesis. I believe that the argument I will present does just that while adequately responding to the claims made earlier in favor of the reality of vat experiences.

To clarify, the purpose here is not to define reality. On the contrary, my hope is to show the unreality of vat experiences while still allowing for the broadest possible understanding of human reality, one that includes the intertwining of human understanding and will, imagination and sense. I want to maintain that everyone experiences and sees the world somewhat differently, but that there is still an out of bounds, so to speak, for reality. In a way then, I am trying to define a limit concept, unreality. My argument only aspires to establish the negative thesis that vat experiences are unreal; it does not assert the reality of any experience. My aim is to locate one boundary that delineates reality from nonreality that will hopefully key in on exactly what makes the experiences of a

brain in a vat unreal.

The argument I shall make is that there exists a *natural* condition to reality. The argument begins with the Berkeleyan distinction between understanding and will that we used above. Recall that I questioned such a strict distinction on the grounds that one cannot perceive the world (i.e. *understand*) without asserting consciously or unconsciously one's unique makeup (i.e. *will*). Nonetheless, human understanding and will are distinct even if they overlap; one does not supervene on the other. Moreover, even though one cannot understand without willing, one can still will without understanding. If one imagines sunbathing on a tropical beach during a philosophy lecture, he incorporates none of the given sense data of the lecture hall. In contrast, watching a person cross the street incorporates both our given sense data of the scene as well as certain aspects of our will—our linguistic constructs, among other things. Thus there is still such a thing as pure will, even if there is no pure understanding. Let us reshape Berkeley's distinction to look like a spectrum, with will on the one side and understanding on the other. Most of the spectrum is a certain combination of both. But on one extreme, there exists pure will which is humanly possible. On the other extreme is pure understanding which is not humanly possible.

To return now to the difference between vat matrices and basic matrices, we now see that a brain in a vat, by definition, never perceives the sense data given to him by his understanding. All the sense data he receives, whether he knows it or not, is the product of the creator's will. Even what he *understands*, is really just the creator's *will*. And much of what he wills – his beliefs, assumptions, past experiences – are the manipulations of a higher will. Even his dreams, hopes and desires are affected by his perception of the world around him. Vat matrices are most akin to the case of imagining oneself on a tropical beach, in that in neither case are we perceiving the sense data that *would have* been given to us had it not been for the computer's will or our own. And indeed, as we said above, the latter case is one of the clearest examples of an experience we usually intuit is not real. In contrast, experiences outside a vat in any of the basic matrices usually represent some intertwining of understanding and will, passively receiving the given sense data and actively interpreting it.

With this distinction now in place, the question becomes: why must experiences that originate from the pure will necessarily be unreal? This is essentially the question we asked above and have put off until now:

Why should reality, or rather, non-reality be contingent upon the source of the experience? The answer, I believe, is that experiences that lie solely within the will do not satisfy the *natural* condition of reality. Beyond the distinction of will and understanding lies the distinction between natural and artificial. *Natural* perception refers to *actively* perceiving (for human perception, contra Berkeley, can never be passive in the sense that one always brings his own beliefs, assumptions, etc. with every act of perception) *the given sense data* in such an honest manner as is appropriate according to oneself or one's species. This is purposefully vague. For example, perhaps a person's reality need not always be guided by the most *literal* or most *probable* construal of sense data. What is important however is that the natural condition of reality limits the will to *some* construal of sense data. It is clear how vat matrices fail this condition. The *natural* condition respects both the given sensations of our understanding as well as the range of appropriate human interpretations of them. Questions like: 'What exactly does this range look like?' and 'When does too much will deem something unreal?' will be questions largely disputed.

Again, to clarify, I am not equating the *natural* with the *real*, but equating the *artificial* with the *unreal*. Reality is constrained by the *natural* but is not defined by it. I think this constraint can be supported by a second look at Ellis. As we said above, Ellis seems to connect reality to the conditions of a particular species. The suggestion we found earlier was that if we are brains in a vat, then our vat worlds should represent the unique reality *for us*. But indeed, those who argue that way entirely miss the point. The point is that the creator of the vat world is interfering with the given, *natural* conditions of the human race—its biology, psychology, philosophy, and rationality among other things. True, Ellis claims that reality varies according to the varying conditions of species. But nowhere does Ellis claim that a species is at liberty to change the given conditions of their species, because there is no absolute reality independent of epistemic values. And if the change is forced upon them, worse off for them. From Ellis' perspective and mine as well, vat worlds hide the true reality of a particular species by interfering and manipulating their *natural* senses.

We began this essay by noting that any brain in a vat matrix presupposes that humans originally existed outside the vat world. I would now add that any vat world implies that humans *naturally* exist outside the vat while the vat world is merely *artificial*. This is so regardless of the

creator of the vat. Moreover, this is so regardless of whether the brain in a vat ever experienced the world outside the vat or if it is even possible to escape. This is true for two reasons. Firstly, should he theoretically escape the vat, say by an act of mercy from the creator, he would be able to exist as before being plugged in. If he cannot still exist outside the vat, then as we said above, a new species evolved with a new reality. Secondly, the creator of the vat is assumed to be someone who did not create man himself, but who manipulated our sense data artificially. I conclude that in any *artificially* created world, experience is not real and the possibility of truth is absent or, at most, discovered accidentally.

In contrast, if we turn now to an analysis of the basic matrices, these matrices all represent *natural* worlds. Suppose, for example, we live in an external world that came into existence as a result of the Big Bang and evolution. The world is a *natural* one. The sensations which present themselves to a person's mind are naturally given to him due to evolutionary forces. Contrast this with the vat worlds discussed above. In those cases, our sensations are imposed upon us not by the whim of nature, but by the whim of some conscious being *over and above nature*. Some may be tempted to say that the vat creators who manipulate our senses are acting according to nature, not over and above it. Yet such a claim misses the point entirely. For just as we would say that a tiger living in the San Diego Zoo is not living in its *natural* habitat, despite its being there through nature's course of events, so too would we say that a machine that puts a human in a vat and re-circuits his brain so that his experiences do not correspond to the material facts they once did, is an action *over and above Nature*.

The Berkeleyan matrix poses an interesting question. As we noted above, it is quite similar to a brain-in-a-vat scenario. What's different is that God is the source of our sense data, not some simulation, and that man has historically never known any other world outside of God's control. How can we explain what is *natural* about such a world? A distinction is in order here, between artificial creation on the one hand and divine creation on the other. I have alluded to this above. God creates the sense data like the creators of vats, but unlike them he also created man. This is crucial. For in the Berkeleyan case, man cannot exist outside the matrix. His being is tied to the matrix world by his very own *creation*, in a story similar to the natural evolutionary one. In a Berkeleyan matrix, then, I argue that experience is real and there is a possibility of truth. The possibility exists because man is *naturally* constituted by those sensations

given to him by God.

To summarize, the *natural-artificial* distinction is crucial for understanding reality and the possibility of truth (or pragmatic truth). For we can agree with Ellis above that truth is human truth. But it is the *natural* abilities of man that are necessary and sufficient for attaining truth. An artificially created vat world denies man of some of his capabilities, interfering with his capacity to attain truth. In an artificially created matrix, there is only an accidental possibility of truth because he is removed from his natural, real world outside it. The artificial matrix hinders his epistemic abilities. In any of the basic matrices, however, either divinely created or not created by any being at all, the possibility of truth is present, for the matrix is indeed the natural world for man.

I will conclude by acknowledging further complications that arise from my suggestion. Where exactly do we draw the line between *natural* and *artificial*? It is natural to dream, so perhaps are my experiences while I am dreaming real? Perhaps man is so *naturally* constituted that he could be said to live in almost two worlds, the world of wakeful states and the world of dream states. Moreover, can drug-induced hallucinations or the hallucinations of schizophrenics be deemed *real* even though they are *natural*? Off hand, I would argue that the dreams and schizoid hallucinations are more *natural* than drug-induced hallucinations, but I acknowledge the problem is still there. Indeed, it may well turn out that the *natural-artificial* distinction is only helpful in demonstrating why vat matrices are unreal, while leaving the question of reality still as wide open as ever.

References

Berkeley, G. (1710). *A Treatise Concerning the Principles of Human Understanding*.

Ellis, B. (1985). *What Science Aims to do*. Images of Science: Essays on Realism and Empiricism with a reply from Bas C. Van Fraassen. Edited by Paul M. Churchland and Clifford A. Hooker. University of Chicago Press

Pryor, J. "What's So Bad About Living in the Matrix?" Retrieved on January 26, 2008 from http://whatisthematrix.warnerbros.com/rl_cmp/new_phil_fr_pryor.html

Van Fraassen, B. (1980). *The Scientific Image*. Oxford. Clarendon Press.