# The Disease Versus Moral-Problem-In-Living Distinction:

## Four Papers in the Philosophy of Medicine

by

REINIER SCHUUR

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

1st March 2024

**ABSTRACT**

The title of my dissertation is 'The Disease Versus Moral-Problem-in-Living Distinction: Four Papers in the Philosophy of Medicine'. This dissertation consists of an introductory chapter and four papers. This dissertation addresses the following question: should medicine be primarily concerned with problems of an involuntary and non-moral nature? This question concerns what I call the 'disease versus moral-problem-in-living' distinction. I do not defend any definition of disease. Rather, I clarify the disease versus moral-problem-in-living distinction, point out the problems that this distinction causes, and argue that we should reject this distinction as a means to constrain the scope of medicine because it denies or undermines the value of moral guidance.

Paper one provides an overview of the concept of disease and mental illness debate. Much of the concept of disease debate has been isolated from developments in philosophy concerning the fact versus value distinction and philosophically methodology. I argue for drawing on such developments to inform the debate on what should be the nature and scope of medicine.

Paper two provides an alternative reading of Thomas Szasz's argument that mental illnesses are not real illnesses. I argue that Szasz's primary concern was not that mental illnesses are not real illnesses. Rather, I argue that his primary concern was the medicalisation of moral-problems-in-living as such, which he saw the concept of mental illness as representing. His solution to this concern was to argue for a clear disease versus moral-problem-in-living distinction to constrain the scope fo medicine. The rest of the dissertation is a criticism of this approach and an attempt to provide an alternative solution to the medicalisation of moral-problems-in-living.

Paper three challenges Gene Heyman's choice model of addiction as a solution to the puzzle of addiction. I argue that Heyman's distinction between two ways of thinking about our best interest, a local versus a global temporal choice framework, is promising in solving the puzzle of addiction. But I argue that Heyman's appeal to a 'natural bias' to the local choice framework has serious problems. I argue that Heyman's assumption that the global choice framework is sufficient for both knowing what is in our best interest and for acting in our best interest gives rise to his appeal to a natural bias as part of his solution. I draw on the work of Hanna Pickard on the role of denial in addiction to argue that Heyman is wrong to assume that the global choice framework is sufficient for knowing what is in our best interest and for acting on such knowledge. I then draw on the work of Gena Gorlin on Cognitive integrity as a framework to explain how we can act contrary to our best interest even when we have knowledge of what is in our best interest.

Paper four clarifies the different ways that the disease versus moral-problem-in-living distinction has been used and then argues against one of those uses. The first is an explanatory distinction within medicine, the second is as a means to constrain the scope of medicine. I make a further distinction between a hard and soft version of the disease versus moral-problem-in-living distinction as a means to constrain the scope of medicine. The hard version of the distinction rests on a narrow conception of disease that has already been rejected in the concept of disease debate. I argue that the soft version of the distinction is incompatible with a broad conception of disease and therefore a broader scope of medicine. I argue that we should therefore reject the disease versus moral-problem-in-living distinction as a means to constrain the scope of medicine.

## ACKNOWLEDGEMENTS

Many people have supported me during the researching and writing of my dissertation. This support has come in many forms: academic, financial, logistical, and moral support.

I want to start by thanking the many supervisors that I have had while working on my dissertation, starting with Hanna Pickard who's work inspired me to get into the philosophy of medicine and psychiatry and who's feedback and advice I continue to benefit from to this day. I want to thank Lisa Bortolotti for her encouragement and support as the supervisor who I started and ended with. I also want to thank Jerry Wakefield, Michael Strevens, and Jim Lennox for supervising me as a Fulbright Scholar at NYU and at Pitt. I want to thank Jim in particular who continues to encourage me and for taking me out on many cycling trips around Pittsburgh. I want to thank Will Davies and Nick Jones for supervising me after my return to Europe from the US. I want to give a special thanks to Iain Law. His supervision helped me to bring my dissertation to its completion.

I also want to acknowledge and thank the many people that have advised and mentored me informally throughout my work on my dissertation. I want to thank Serife Tekin both for her role as editor of my first published book chapter and for her moral support of me as a young scholar. I want to thank both Dale Stevens and Corinne Bloch-Mullins for mentoring me over the years on how to navigate both the academic and non-academic aspects of scholarly life. I want to especially thank Greg Salmieri and Gena Gorlin both for making my visit and eventual return to the United States possible.

I want to thank the Fulbright Commission of the Netherlands and Elsevier for co-sponsoring and funding me as a visiting Fulbright Scholar at NYU and at the University of Pittsburgh. I also want to thank the John Tempelton Foundation for funding my travels and accommodations for the Genetics and Human Agency conferences.

Finally, I want to thank my friends and family both for their feedback and support. I want to thank Caspar Safarlou, Shahin Kaveh, Nino van Staaden, and Kritika Maheshwari for their helpful feedback on my dissertation papers. I want to thank all of my siblings for their moral and social support over the years. I want to thank my Mother, Gea Visscher, both for her financial support and especially for the many years of encouragement to pursue academic achievement. I want to thank my dear wife, Tolani Olonisakin, for inspiring me to persist and grow in our shared love of learning. I want to thank my Father, Siebe Schuur, who deeply regretted before his passing that he would not see me finish my dissertation and be awarded my PhD degree. His love of knowledge has served as an indispensable encouragement both for this dissertation and beyond. Thank you.

TABLE OF CONTENTS

Abstract i

Acknowledgements ii

**Introduction** 1

**Mental Health and Illness: Past Debates and Future Directions** 11

—1.0 Origins, assumptions, and main positions 12

1.1 Thomas Szasz's conceptual analysis 13

1.2 Normativism 15

1.3 Naturalism 16

1.4 Hybrid accounts 17

—2.0 Future directions 18

2.1 Traditional conceptual analysis 18

2.2 Facts and values 19

Conclusion 21

References 22

**Szasz and the Disease Versus Choice Distinction: an Unanswered Question for Medicine** 23

1. Thomas Szasz's functional analysis of the concept of mental illness 24

2. Thomas Szasz's conceptual analysis: real illnesses as biological and value-free 27

3. Thomas Szasz and the disease versus choice distinction 31

4. The disease versus choice distinction as an unanswered question for medicine 33

References 36

**The Puzzle of Addiction, Knowledge of the Good, and Cognitive Integrity** 40

—1.0 Heyman's analysis of and solution to the puzzle of addiction 43

1.1 Heyman's analysis of the Orthodox view of addiction 44

1.2 Heyman's criticism of the Orthodox view of addiction 45

1.3 Heyman's solution: voluntary behaviour as biased towards self-destruction 49

—2.0 Criticism of Heyman's analysis of and solution to the puzzle of addiction 54

2.1 Criticism of Heyman's analysis of the Orthodox view 55

2.2 Criticism of Heyman's solution to the puzzle of addiction 61

2.3 Heyman's assumption about voluntary behaviour made explicit 63

—3.0 Cognitive integrity as an alternative framework for solving the puzzle of addiction 64

3.1 Hanna Pickard's work on denial as a solution to the puzzle of addiction 65

3.2 Cognitive integrity as an epistemic virtue and its role in knowledge acquisition and action 67

3.3 Cognitive integrity as a framework for solving the puzzle of addiction 69

Conclusion 71

References 73

---

**Rejecting the Disease Versus Moral-Problem-in-Living Distinction to Constrain Medicine** 77

—1.0 The disease versus moral-problem-in-living distinction 79

1.1 The disease versus moral-problem-in-living distinction as explanatory 81

1.2 The disease versus moral-problem-in-living distinction to constrain medicine 86

1.3 The costs of the disease versus moral-problems-in-living distinction to constrain medicine 88

—2.0 The hard and soft versions of the disease versus moral-problem-in-living distinction 91

2.1 Criticism of the hard version of the disease versus moral-problem-in-living distinction 91

2.2 The soft disease versus moral-problem-in-living distinction to constrain medicine 96

—3.0 Rejecting the soft disease versus moral-problem-in-living distinction as demarcation 99

3.1 Why we should reject the soft disease versus moral-problem-in-living distinction 100

3.2 The soft disease versus moral-problem-in-living distinction, a harmful trade-off 103

3.3 Addressing the motivations behind the disease versus moral-problem-in-living distinction 105

Conclusion 107

References 109

**Introduction**

The following four papers are a dissertation by papers that broadly falls under the topic of the problem of defining the nature and scope of modern medicine. This topic is sometimes referred to as 'the problem of demarcation', 'the line drawing problem' or 'the concept of disease debate'. Such debates tend to focus either on how to distinguish 'medical' from 'normal' suffering or on the role of non-biological models and conceptions of diseases and medical disorders, especially within the context of psychiatry and mental illness. This focus is reflected in the fact that 'medicalisation' is primarily understood as either 'pathologisation', i.e., regarding some condition as abnormal suffering, or 'biologisation', i.e., regarding some condition as distinctly biological.

But while some of the concept of disease debate is about whether a particular condition is a disease or concerns the role of biology in medicine, much of what animates the whole concept of disease debate historically are questions regarding the nature and role of medicine in human life as such. Indeed, starting in the 1960's, there were criticisms that modern medicine and psychiatry were being extended and misapplied to conditions that should not be medicalised. These criticisms were also arising during a time that many of the misapplications of the concept of disease in the past were being studied. Some of theses criticisms went beyond criticisms of concrete misapplications, arguing that even the medical conditions that we regard as uncontroversial are nothing more than a reflection of medicine's and society's values. Indeed, 'medicalisation' is very often regarded as an undesirable phenomenon that must be contained.

A large part of the motivation behind the clarification of the concept of disease was therefore to argue that modern medicine is justified in regarding many condition as diseases on non-arbitrary grounds, and to argue that medicine is therefore a legitimate institution. Clarifying the concept of disease was in large part a reactionary project meant to legitimise medicine in the face of such criticisms, instead of the concept of disease being proactively developed to guide medicine and clarify the relationship between health, disease, and other problems in living.

The concept of disease debate as a debate about the nature, legitimacy, and scope of medical science and practice has tended to focus on four major issues in defining diseases: (1) facts versus values, (2) biology versus non-biology, (3) medical versus normal suffering, (4) and pathology versus mere difference. One criticism of the concept of disease was that it primarily or only reflects the values of society and does not refer to any 'value-neutral' facts. One response to this kind of criticism has been to argue that there are certain kinds of value-neutral facts, like increased mortality, that distinguish diseases from other things we don't like. Such accounts, however, were criticised for failing to explain how the concept of disease does and should reflect the values of patients as individuals. Much of the debate about the concept of disease has since focused on the relationship between facts and values in defining disease as a means to address various criticisms regarding the nature and scope of medicine. The other three issues can in a certain sense be regarded as different ways in which this issue comes up. The issue of biological versus non-biological models in medicine concerns which kinds of scientific facts are relevant to defining disease. The issue of medical versus normal suffering and pathology versus mere difference concern which facts and values should constrain what are legitimate diseases.

The theme with which these papers are concerned is a distinct issue in the concept of disease debate that has largely been neglected by the literature. This issue is distinct from the role of facts and values in defining the concept of disease. It is the issue of whether or not (and subsequently to what extent) we should conceive of diseases as involuntary and non-moral in nature (and by extension whether or not and to what extent problems of a voluntary and moral nature belong under the purview of medicine). Another way to understand this issue is to contrast it with the different conceptions of what 'medicalisation' entails. Medicalisation often either means the process of pathologising a condition, i.e., marking it out as a medical as opposed to a normal form of suffering, or it means the processes of biologising a condition, i.e., regarding a problem as primarily biological in nature as opposed to social, for example. But the concept of medicalisation has also been used to imply the denial of agency or of the moral nature involved in certain conditions. This kind of medicalisation may be termed as a 'pacification' or 'demoralisation'. This kind of medicalisation is more specific than the other two, since it raises the question of whether problems of a voluntary and moral nature should fall under the purview of medicine. I call the attempt to constrain medicine to non-moral problems *the medicine versus morality distinction*.

I first came across this distinction implicitly in debates around the nature of addiction, where some argued that addiction is not a moral problem but a disease because addictive behaviour is involuntary. But I also came across explicit defences of the medicine versus morality distinction as such, in particular in the work of Thomas Szasz (Szasz, 1961). Indeed, much of my work in the four papers that follow are a response to Szasz's argument that medicine should be constrained to involuntary conditions in order to prevent the medicalisation of voluntary and moral-problems-in-living. But I am not just responding to Szasz. I argue that Szasz's position, that some kind of *disease versus moral-problem-in-living distinction* needs to be made in order to constrain the scope of medicine, is actually the default position. I argue that the disease versus moral-problem-in-living distinction should not be used to constrain the scope of medicine. Rather, I argue that this distinction should only be used *within* the context of medicine to distinguish between different kinds of conditions. I argue that we should reject the disease versus moral-problem-in-living distinction to constrain the scope of medicine because it undermines the role of choice and moral-problems-in-living in health related issues. Contrary to Thomas Szasz, I argue that distinctly moral-problems-in-living should fall under the purview of modern medicine.

This dissertation is novel in that it explicitly challenges the medicine versus morality distinction on its own terms rather than in relation to a specific set of conditions. While only the fourth paper challenges the medicine versus morality distinction directly, the other papers all address specific issues that are relevant in challenging this distinction. Moreover, the papers deal with a whole range of issues, such as the legitimacy of medicine, agency, stigma, and the underlying assumptions that frame the way we think about medicalisation and what role medicine should play in our lives. Due to the large array of issues that the medicine versus morality distinction raises, not every one of them can be resolved within the scope of these four papers. Instead, the purpose of this dissertation is to put these issues into a wider context in order to make a compelling case for why we should reconsider the medicine versus morality distinction.

Paper one is based on an earlier version of 'Mental Health and Illness: Past Debates and Future Directions', my chapter in *The Bloomsbury Companion to Philosophy of Psychiatry*. The purpose of this paper in this dissertation is to bring developments in mainstream academic philosophy to bear on the concept of (mental) illness debate and to use such developments to argue for a specific way that we should reframe and pursue this debate into the future.

I argue in this paper that the concept of illness debate is only marginally about where the line between health and disease should be. The debate is fundamentally about the basis of the concepts of health and illness (and by extension about the legitimacy of medicine in general and psychiatry in particular). Indeed, I frame the concept of illness debate as primarily a reaction to Thomas Szasz's criticism of the concept of mental illness, which spurred on a deeper debate about the concepts of health and illness and the legitimacy of medicine. In particular, what is really at issue in the concept of illness debate is the relationship between facts and values in constraining the concept of illness and by extension the role of medicine. Different positions in this debate, Naturalism, Normativism, and Hybrid accounts, are different accounts of how we should think about the relationship between facts and values in defining disease and the role of medicine in our lives. But I argue that the methodology by which such relationships are unearthed (conceptual analysis) and whether it primarily seeks to capture what we mean by disease versus what diseases are, is unclear. For those that argue that medicine is fundamentally legitimate, conceptual analysis both captures what people mean by the concept of disease and what diseases are. But those who are more critical of medicine's authority use conceptual analysis to argue that there is a discrepancy between what we say diseases are (that they have a value-free basis) and what they actually are (that they are not value-free but reflect the values of society).

I argue in this first paper that there have been significant developments in mainstream academic philosophy regarding the both fact versus value distinction and the role of conceptual analysis. I argue that much of philosophy has moved beyond the fact versus value distinction, and the role of conceptual analysis has changed. In particular, conceptual analysis has moved beyond its reliance on what I call *conceptual conservatism*, the view that our intuitions must be balanced in the way that we develop and define our concepts. While intuitions can be helpful in determining the bounds of conceptual analyses, I argue that developing a concept may also violate at least some of our intuitions if doing so helps to develop the usefulness of the concept in question.

While following papers in this dissertation mention the fact versus value distinction and issues of philosophical methodology, none of them attempt to resolve those issues beyond what is mentioned in this first paper. In the past, as will be mentioned in the first paper, I regarded the fact versus value distinction as crucial to responding to Szasz. But this is no longer the case. In paper two, for example, I argue that Szasz's claim that mental illnesses are not real illnesses is not based on the fact versus value distinction but rather on the disease versus choice distinction, which is a distinction that one can make independently of the fact versus value distinction. I do, however, rely in following papers on the rejection of the project of balancing our intuitions when thinking about the nature and scope of medicine. Indeed, Szasz's defence of the medicine versus morality distinction is in many ways motivated to preserve our intuitions. My project intends to sacrifice some of our intuitions so that we can rethink the nature and scope of medicine.

Paper two, ‘Szasz and the Disease Versus Choice Distinction: an Unanswered Question for Medicine’, has three distinct goals. First, I argue that Thomas Szasz’s thesis that mental illnesses are not real illnesses has been widely misunderstood and I offer an alternative way to read his argument for this claim. Second, I criticise Thomas Szasz’s thesis that mental illnesses are not real illnesses based on my reading of this claim. Third, I argue that Thomas Szasz’s claim that mental illnesses are not real illnesses is based on a wider claim that presents an ‘unanswered question’ for modern medicine. The first goal involves laying out Szasz’s two ‘central claims’ that mental illnesses are not real illnesses: (1) that the concept of mental illness functions to deny agency both in individuals and to undermine the very notion of agency itself, especially agency regarding moral-problems-in-living and (2) that mental illnesses are not real illnesses because of the meaning of the concepts of illness and mental illness. I point out that, for Szasz, the claim that mental illnesses are not real illnesses is an analytical truth, that mental illnesses cannot be real illnesses because of the meaning of these two concepts. I argue that because Szasz’s argument is primarily motivated by the agency-denying function of the concept of mental illness that any reading of Szasz’s thesis requires explaining how this agency-denying function of the concept of mental illness relates to his claims about the meaning of the concepts of illness and mental illness. I then present the two main ways that Szasz’s claim that mental illnesses are not real illnesses has been read: (1) that mental illnesses are not real illnesses because real illnesses are deviations from biological norms (or have lesions) and mental illnesses are deviations from non-biological norms (or lack lesions): (2) that mental illnesses are not real illnesses because real illnesses are deviations from objective norms (i.e., facts) whereas mental illnesses are deviations from subjective norms (i.e., values). I argue that neither of these readings can explain how Szasz’s two central claims are related. Szasz’s claim about lesions concerns his claim that traditionally illnesses were defined by some kind of distress and the presence of a lesion in order to distinguish medical illnesses from normal suffering. Szasz claims that the concept of mental illness is a rejection of the lesion criterion. I argue, however, that these meanings Szasz ascribes do not explain why he takes his thesis to be an analytical claim, as the meanings are not mutually exclusive, and they do not explain the agency-denying function of the concept of mental illness.

Szasz attributed many meanings to both the concept of illness and mental illness, but it is not ‘distress + lesions’ and ‘any distress' that can explain his two central claims. Moreover, Szasz’s claim that mental illnesses are not real illnesses was also directed at non-biological views of mental illness, such as psychoanalysis and behaviourism. I also argue that reading Szasz as claiming that mental illnesses are not real illnesses because real illnesses are deviations from objective norms whereas mental illnesses are deviations from subjective norms is a misreading of his position. This reading runs contrary to my own view in the past that Szasz’s claim that mental illnesses are not real illnesses was that bodily illnesses are based on facts and mental illnesses are based on values. I no longer read Szasz’s position this way. Szasz himself rejects this reading. Furthermore, this reading cannot explain the agency-denying function of the concept of mental illness since one can deny agency regardless of whether the norms are objective or subjective. Szasz instead argues that real illnesses are deviations from norms of how the body ought to function, whereas mental illnesses are deviations from norms of how persons ought to be and act.

The latter points to another set of meanings that Szasz attributes to the concepts of illness and mental illness, namely, that illnesses are involuntary in nature whereas mental illnesses are voluntary and therefore moral in nature. 'Moral' here does not mean merely normative but those norms relevant to evaluating and guiding actions and other aspects of the self we are responsible for, such as our character traits, our very selves, and how we use our minds. That is, Szasz regards the meaning of the concept of illness as implying an involuntary condition that we are therefore not directly morally responsible for, as opposed to things we are morally responsible for such as our actions and our character. Szasz regards the concept of mental illnesses to mean a learned behaviour and therefore a condition that is not involuntary but involves voluntary and therefore moral components that are subject to moral evaluation and guidance. I argue that Szasz has these two meanings of the concepts of illness and mental illness in mind when he argues that his claim that 'mental illnesses are not real illnesses' is an analytical claim.

Furthermore, it is this meaning that explains why Szasz thinks that the concept of mental illness has the agency-denying-function it does, namely, that the claim that mental illnesses are real illnesses is to claim that conditions that are voluntary are involuntary. In other words, what motivates Szasz's claim that mental illnesses are not real illnesses is the view that medicine ought to be constrained by a disease versus choice distinction where medicine is primarily concerned with conditions of an involuntary nature. Szasz's view is therefore that the claim that 'mental illnesses are real illnesses' implies a rejection of the disease versus choice distinction, and by extension a rejection of the medicine versus morality distinction. Szasz therefore regards the concept of mental illness as implying the medicalisation of moral-problems-in-living as such.

Subsequently, I criticise Szasz's position. First, by arguing that the claim that mental illnesses are real illnesses is (more often than not) a claim that mental illnesses involve medically relevant dysfunction and suffering. Second, by arguing that if what people truly mean by the claim that 'mental illnesses are real illnesses' is that mental illnesses are involuntary that this claim is therefore precisely not analytically false. So even if what people primarily mean by an illness is an involuntary condition and they view mental illnesses as real illnesses, then all Szasz can claim is that such a claim is empirically false, not that such a claim is analytically false. But while Szasz's claim that mental illnesses are not real illnesses does not follow, the basis of this claim raises an unanswered question. Namely, do we really need a disease versus moral problem-in-living distinction to constrain the scope of medicine? Even if Szasz is wrong about modern medicine rejecting such a distinction to constrain the scope of medicine, the philosophical question of whether (and if so, how) a disease versus moral-problem-in-living distinction should be used to constrain medicine remains unanswered. That is, the question of whether (and if so, how) medicine should relate to moral-problems-in-living remains unclear and poses an unanswered question of how we should think about the nature and scope of medicine. Papers three and four are in many ways an answer to this question. Paper three deals with the question of how to make sense of the way choice and knowledge of the good can explain certain conditions in medicine. In this case, how voluntary and moral explanations could resolve the puzzle of addiction. Paper four argues that we should not use the disease versus moral-problems-in-living distinction to constrain to scope of medicine because doing so undermines the value of moral guidance within medicine.

Paper three ‘The Puzzle of Addiction, knowledge of the good, and cognitive integrity’ is based on my winning 2022 Peter Sowerby Contest essay titled ‘The Puzzle of Addiction: knowledge of the good as an achievement’. The purpose of Paper three is to present an alternative way to think about the puzzle of addiction. In particular, to present a view that explains how we can freely act contrary to our own good even when we know what is in our own good.

The puzzle of addiction is the puzzle of explaining why addicts keep using drugs despite the severe negative consequences. Addictive behaviour is supposed to be a puzzle because of the assumption that, if our actions are voluntary and the costs of our actions outweigh their benefits, that we would stop engaging in such behaviour. But this doesn’t appear to happen in addiction. The ‘Orthodox view’ of addiction solves this puzzle by arguing that addicts keep using drugs because addictive behaviour is somehow involuntary and therefore a disease. However, it has been argued in recent decades that addictive behaviour is not involuntary but voluntary. This conclusion again raises the puzzle of addiction of why addicts keep using drugs. Gene Heyman has argued that we need to rethink our assumptions about the nature of voluntary behaviour in order to explain how harmful voluntary behaviour is even possible. In particular, Heyman argues that the assumption that voluntary behaviour is necessarily rational with regards to acting in our best interest gives rise to the view that harmful behaviour, like addictive behaviour, is necessarily involuntary behaviour. Heyman’s solution to the puzzle of addiction is to argue that voluntary behaviour is therefore not necessarily rational with respect to acting in our best interest, but that voluntary behaviour is ‘naturally biased’ to irrationality with respect to acting in our best interest.

The purpose of this paper is to build on aspects of Heyman's solution that I think are correct and to complete what I think his solution is missing. I agree that Heyman’s work on the distinction between temporal frameworks is a valuable part of the solution to the puzzle of addiction. However, I argue that the way in which he conceives of how these frameworks relate to knowledge of our best interest results in his appeal to a natural bias, which I argue has flaws. I also argue that Heyman’s analysis of what causes the puzzle of addiction as well as his analysis of the existing solutions are also flawed, and that part of his solution shares many of those flaws.

Heyman’s view is that what causes the puzzle of addiction is the view that voluntary behaviour is necessarily rational with respect to acting in our best interest, a view I call ‘homo economicus’, and he argues that this view leads to the Orthodox view that addictive behaviour is necessarily involuntary. He argues that both homo economicus and the Orthodox view are false by showing that addictive behaviour is not involuntary but is in fact voluntary. In response, I argue that Heyman has misread the Orthodox view of addiction. I argue that there are two versions of the Orthodox view, one that does regard addictive behaviour as involuntary in the way that Heyman defines it, and another version that defines addictive behaviour as involuntary in a broader sense. Neither of these views, I argue, holds that addictive behaviour is involuntary because voluntary behaviour is necessarily rational with respect to our best interest. Instead, I argue that the Orthodox views are more closely related to Heyman’s own view, the view that voluntary behaviour is biased towards irrationality with respect to our best interest, a view that I call ‘homo irrationalis’. While Heyman presents his own solution as being contrary to that of the Orthodox views, I argue that they actually have a lot more in common than it first appears.

Heyman's solution to the puzzle of addiction comes in two parts. First, he distinguishes between the 'local' and 'global' choice frameworks. These are distinct temporal frameworks from which agents calculate what is in their best interest from either a short term (local) or a long term (global) perspective. Heyman argues that, from the local choice framework, many of the actions of addicts are in their best interest and that it is only from the global choice framework that their actions appear to be contrary to their best interest. However, Heyman takes the global choice framework to be sufficient for knowing what is in one's best interest and for acting on such knowledge. This creates a problem for Heyman. Namely, why do we so often act contrary to our best interest if framing our actions from the global choice framework is better for our overall best interest? In response, Heyman argues that we all have a natural bias to frame the costs and benefits of our actions from the local choice framework, i.e., to act irrationally with respect to our best interest from the global choice framework. This is the second part of Heyman's solution.

I criticise the second part of Heyman's solution, his appeal to a natural bias, because it is ill-defined and thereby loses much of its explanatory power. In particular, I argue that the natural bias solution shares a lot in common with the weak Orthodox view in that the latter also has a vague notion of involuntary behaviour that is ill-defined therefore lacks explanatory power. In this sense, I argue, Heyman's natural bias solution is not a major advancement over the weak Orthodox view that he intends to reject and replace. I argue that what is really at issue is the underlying assumption (which is shared with both homo economicus and homo irrationalis) that something like the global choice framework ought to be sufficient for acting in our best interest. The disagreements lie in whether we are fully free to access to such a framework. It is this assumption that gives rise to Heyman's appeal to a natural bias to the local choice framework.

The primary purpose of this paper is to question this assumption and to provide an alternative framework for solving the puzzle of addiction. First, I draw on Hanna Pickard's work on the role of denial in addiction. Pickard argues that gaining knowledge of the causal relation between our actions and their negative consequences is not obvious, but actually a cognitive achievement. I argue therefore that merely taking a global perspective on ourselves is not sufficient for knowing what is in our best interest, nor for which actions are required to achieve it. Denial can also explain why some addicts keep using drugs despite the negative consequences. But denial cannot fully solve the puzzle of addiction since acceptance of knowledge that continued drug use is bad for you is rarely sufficient for addicts to stop using drugs. Instead I draw on the work by Gena Gorlin on Cognitive integrity to provide an alternative framework for solving the puzzle of addiction. Cognitive integrity refers to the epistemic virtue of a willingness to gain knowledge and to know what is true (which also requires taking a long term perspective on one's self). However, such a virtue requires constant choice to develop and maintain. One can therefore fail to act in one's best interest despite having knowledge of what is in one's best interest because acting on knowledge of what is in one's best interest requires choosing to bring to awareness the full context of such knowledge. The global choice framework is therefore not sufficient to know what is in one's best interest or to act in one's best interest because both requires of us to maintain such knowledge while acting. This solution to the puzzle of addiction draws on the global choice framework without any of the problems of the 'natural bias' solution.

Paper four, titled ‘Rejecting the Disease versus Moral-Problem-in-Living Distinction to Constrain Medicine’, is about the medicine versus morality distinction and seeks to challenge it. The purpose of this paper is to argue that we should not use the disease versus moral-problem-in-living distinction to constrain the scope of medicine to primarily problems of an involuntary and non-moral nature. I define moral-problems-of-living for the purposes of this paper as those problems of the self that are subject to voluntary and moral guidance. It is not the purpose of this paper to argue how exactly we should distinguish between conditions which are subject to moral guidance and those which are not. The purpose of this paper is to argue that we should develop such a distinction only to separate different kinds of explanations within the scope of medicine (rather than to use such a distinction to constrain the scope of medicine to non-moral conditions).

In Part one of Paper four, I explain how the disease versus moral-problems-in-living distinction first developed as a distinction between different kinds of explanations at a time when the boundaries between medicine and other domains such as morality were less clear and more fluid. Indeed, moral explanations were often used for certain medical conditions, especially in psychiatry. Such moral explanations were not necessarily about blame, but about understanding how those aspects of the self that are under our voluntary control require normative guidance. That is, normative guidance with respect to what kind of selves and traits we need to develop in order to properly function with regard to our selves and actions. Such moral explanations started to be removed from the domain of medicine for various reasons, such as physicians wanting to push out or demote the role of non-scientifically trained lay healers in medicine. But the reasons for removing such moral explanations from medicine that are still relevant today are twofold. First, the misapplication of narrowly medical explanations and treatments to problems of a voluntary and moral nature, i.e., moral-problems-in-living. Second, the misapplication of narrowly moral explanations and approaches to problems of an involuntary and non-moral nature. In other words, the main problem at issue here are concerns over the medicalisation of moral-problems-in-living and the moralisations of medical conditions. Such misapplications would indeed be a problem in so far as both the conditions, and the explanations and treatments for them, are narrowly construed. Why? If medicalisation does imply a denial or reduction of the patient’s agency, then the medicalisation of a condition that in fact involves more agency than we think, would be a denial or reduction of such agency. A similar concern arises in the moralisation of a condition that is in fact an involuntary and thereby non-moral problem. The particular concerns regarding such misapplications are that wrongly moralising a condition can lead to blaming and stigmatising a patient, while wrongly medicalising a moral-problem-in-living can disempower a patient. There is also the concern that our very conception of ourselves as agents facing distinctly moral-problems-in-living are threatened by medicalisation if the domain of medicine is not constrained to involuntary and non-moral conditions. The solution to such misapplications has been to define medicine as primarily concerned with problems of an involuntary and non-moral nature. And therefore, the solution has included the usage of the disease versus moral-problems-in-living distinction to separate medicine and morality. But I argue that this solution also has negative consequences, such as undermining voluntary and moral explanations in medicine because such explanations would now undermine the medical status of conditions subject to such explanations.

In Part two of Paper four I explain two versions of strategy mentioned above to prevent the problems of misapplication, the hard and the soft version. The hard version seeks to entirely separate diseases as involuntary and non-moral conditions from voluntary and moral-problems in living. Examples of those who advocate for such a hard version of this distinction are Thomas Szasz as well as early critics of the inclusion of moral conditions and explanations within medicine. But this hard version of the disease versus moral-problems-in-living distinction to constrain the scope of medicine has been rejected in virtue of the larger debate about the concept of disease in the 1960's. The consensus that came out of that debate is that disease should not be defined by a particular causal process, such as an involuntary one. Instead, the consensus became that diseases should be defined primarily by their negative medical consequences. The main debate that ensued was on how to distinguish medically relevant negative consequences from medically irrelevant negative consequences. But the rejection of the idea that diseases should be involuntary opened up the possibility that diseases could involve voluntary and moral components. But the rejection of the idea that medicine should conceive of diseases as primarily involuntary and non-moral did not imply a rejection of the medicine versus morality distinction. Indeed, all of the concerns about the misapplication of medical and moral categories which motivated that project remained. The solution to this problem has been to argue instead for a soft version of the disease versus moral-problem-in-living distinction to constrain the scope of medicine and to prevent problems of misapplication. The soft version of this distinction, as articulated and defended by scholars such as John Sadler, holds that conditions may have significant moral components if they also meet the general criteria for being a disease. However, those general criteria are often regarded as being inherently non-moral in nature, such as an underlying dysfunction. On this view, then, the general criteria for disease will not allow voluntary and moral components to disqualify a condition from being a disease, while still allowing the usage of such criteria to distinguish between diseases and 'purely' moral-problems-in-living. The soft version of the disease versus moral-problems-in-living distinction thereby attempts to prevent problems of misapplication while not creating incentives against moral explanations in medicine.

In Part three of Paper four I argue that we should reject the use of the soft version of the disease versus moral-problem-in-living distinction to constrain the scope of medicine as a strategy to prevent problems of misapplication. First, I argue that the soft version of the distinction presumes that the general criteria for disease qualification are inherently non-moral in nature. The primary candidate for such a 'non-moral feature' that has been proposed are dysfunctions. However, dysfunctions can be moral in nature if they involve a failure of a function that is voluntary or a learned process subject to moral guidance, such as one's behaviour or moral emotions. The other criterion that has been proposed to limit moralisation in medicine is that conditions should be involuntary. But this criterion is incompatible with the view that diseases should not be defined as being involuntary but rather be primarily defined by their negative consequences. I argue that the soft version of the disease versus moral-problem-in-living distinction necessarily undermines voluntary and moral explanations in medicine. Moreover, I argue that concerns of misapplications are better addressed by the further development the distinction between involuntary or voluntary and non-moral and moral processes and conditions within the scope of medicine.

This dissertation is motivated by a concern that I share with Thomas Szasz that modern medicine can often be in tension with our conception of ourselves as moral agents. But quite unlike Szasz I argue that the solution to this should not be to maintain a disease versus moral-problem-in-living distinction to constrain the scope of medicine. Rather, my solution is to reject this distinction and instead argue for an integration of medicine and morality where relevant. Both modern medicine and morality are too valuable in their potential for human life to separate.

This dissertation is also motivated by my concern that medicalisation is too often regarded as an undesirable consequence that needs to be contained. Relatedly, I am concerned that much of the concept of disease debate has been too reactive in its response to prevent the harms of medicalisation by its strategy of constraining the scope of medicine. While I acknowledge that medicalisation can be harmful in many cases, my own perspective is that constraining the scope of medicine is not the only way to respond to such harms, and indeed, that such a response itself can have negative consequences. My own strategy, reflected in this dissertation, is to address concerns of both medicalisation and moralisation not primarily by constaining either of them but by reforming what the domains of medicine and morality entail. My primary concern with medicalisation is therefore not that the domain of medicine is ever expanding, which is one of the main concerns motivating the concept of disease debate. My primary concern with medicalisation is what medicalisation, and the nature of medicine itself, entails. To be more specific, I am primarily concerned that medicalisation can and often does imply a denial or an undermining of the importance of the agency of patients. While I regard it as very important to figure out what patients do and do not have control over, my preference is to separate such questions from the process of medicalisation as such. Rather, under my own conception, medicalisation should entail the application of medical knowledge and therapies to any and all human problems of living. Much of the debate regarding such applications of medicine I would want to focus on whether or not particular explanations and therapies are appropriate for a condition or even required at all.

I recognise that my approach to concerns of medicalisation stands in contrast with other scholars. It is also beyond the scope of this dissertation to elaborate on, let alone defend, my wider perspective on medicalisation and the nature and scope of medicine. But I want to make explicit my broader motivation and perspective on medicalisation in order to make clear where this dissertation stands in relationship to my thinking on this topic. I have delimited myself in this dissertation to addressing a particular concern regarding medicalisation, namely, the concern that the expansion of medicine poses a threat to problems of a voluntary and moral nature being improperly explained as involuntary and treated as if they are non-moral in nature. Indeed, I argue that this was Thomas Szasz's primary concern regarding the process of medicalisation.

What I have tried to do in this dissertation is to argue that this concern regarding the medicalisation of moral-problems-in-living should not be addressed with the strategy of removing such problems from the domain of medicine. Rather, the strategy that I have argued for is to reform what medicalisation entails by allowing for moral-problems-in-living to fall under the purview of medicine, and to instead argue that our attention should be focused on better distinguishing between things we do and do not have control over and problems that are and are not subject to moral guidance. This strategy calls for reforming medicalisation, not constraining it.

**Mental Health and Illness: Past Debates and Future Directions**

Psychiatry is often taken to represent the application of medical science and practice to the human mind. Modern medicine has clearly done a lot to alleviate all manner of suffering from physical ailments that were once a normal part of life[1]. It has also helped reduce the stigma of certain physical conditions that were once thought to be signs of a weak moral character for which one is responsible. But what exactly the application of medicine to the realm of the human mind entails is not always so clear. To the extent that medicine concerns itself with the treatment of illness, it implies that the mind can be ill in some sense, that there can be such a thing as *mental* illnesses. Psychiatry is therefore immediately confronted by questions of a conceptual nature. What exactly would it even *mean* for the mind to be 'ill'? How do mental illnesses relate to other problems that we face in life, such as struggling with loss, or perhaps our own character flaws? How does the concept of mental illness fit in with our conception of ourselves as free agents who make decisions for which we are responsible? These are questions of philosophy, which deal with clarifying how we *conceive* of ourselves and our place in the world, and *develop*ing our understanding of this fundamental relationship. Psychiatry in particular raises the following questions: do we have any psychological capacity to *change* ourselves, what would that even mean, what would constitute a change for the better, and what should we not change but accept as healthy forms of suffering, like grief? If psychiatry is the application of medicine to the human mind, one concern is that we will address these questions of who we are and how we should live our lives with narrowly biomedical answers, that is, that we will attempt to alleviate all forms of suffering, such as grief, as diseases to be treated, and that we will deny any and all failures of character and responsibility. This would threaten our self-conception as free agents who have responsibility for who we are and who can change ourselves for the better, and the idea that some struggles are a normal or even be a valuable part of life. This raises a question that has been the subject of much debate. What do various medical concepts, such as health, disease, and illness, even mean in general? And further, are we justified or not in applying those concepts to the things that we call mental illnesses in mental health sciences and practices?

This paper presents the past answers to this question and the future directions for this debate. In part one, I present the origins of this debate with Thomas Szasz's criticism of psychiatry and his argument that mental illness is a myth. This includes explaining two widely held assumptions in this literature; the validity of the traditional philosophical method of conceptual analysis and the fact/value distinction. In part two, I present the three reactions to Szasz's conceptual analysis: Naturalism, Normativism, and Hybrid accounts. In part three, I show how the two aforementioned mentioned assumptions of the existing literature can no longer be taken for granted given the current state of philosophy, and that taking these developments seriously can give us a deeper and different perspective on the issue of applying medical concepts to problems of the mind.

---

[1] Tooth decay used to be rife, now it is a disease. Improving standards changed what it means to be healthy.

## —1.0 Origins, Assumptions, and Main Positions

The debate about the meaning of medical concepts originated with Thomas Szasz's criticism of psychiatry in the 1960's[2] (Szasz, 1961). He argued that psychiatry not only is used as a tool for social-control[3], but that it represents the misapplication of medical concepts to what he called normal 'problems of living'. Psychiatry certainly does face a lot of criticism that it medicalises 'normal suffering'[4]. But Szasz held that most, if not all, of the conditions that psychiatry treats are problems of living, such as schizophrenia. This position implies that the concept of 'mental illness' does not refer to actual illnesses at all: mental illness is a myth. However, Szasz did not deny that people had schizophrenia or could suffer from it. What, then, did he mean? What did Szasz take the concept of illness in general to mean such that he rejected its application to the mind?

The deeper question that what we should really care about here is, how we can ascertain the meaning of the concept of illness in order to settle which conditions to apply the concept to? The subsequent question of how we should determine the meaning of concepts is an issue of philosophical methodology. There are many assumptions about the nature of concepts in the concept of illness debate. For instance, many contributors to this debate in the past assume that concepts have necessary and sufficient conditions or that their meaning is accessible *a priori*, that is, prior to experience and accessible through reflection alone. But there are always exceptions to such specific assumptions. A more general assumption in the concept of illness debate appears to be some kind of *conceptual conservatism*, that is, the view that the way in which we generally use concepts has prescriptive import for how we should use concepts. This view contrasts with the view that analysing the meaning of the concept only gives us linguistic meaning, but that it is entirely possible that we should scrap or modify the meaning of the concept for a better alternative. This conceptual conservatism in the concept of illness debate is reflected in the respect that is given to intuitions both as a source of the meaning of a concept and as a test for whether and how we should use a concept. For example, if a definition of disease considers ageing to be a disease, then this violates our intuitions since we generally do not regard ageing as a disease. Such a violation of our intuitions is given as evidence that the definition in question fails both as a description of how we *do* use the concept of disease and as evidence that we should not use such a concept of disease. Any debate about the nature of disease has to start with a generic conceptual analysis of what we happen to mean by the concept of disease, even if we end up radically changing it. But what I call traditional conceptual analysis holds that the way in which we use such concepts also tells us *how* we should use them. People engaged in traditional conceptual analysis differ on which intuitions they want to preserve, but they agree that we should preserve either the most or our deepest intuitions about concepts.

---

[2] Szasz is often labelled as being a part of the anti-psychiatry movement, a label he rejected, which was a reaction to the history of involuntary treatment and other abuses by psychiatry. Other figures include Michel Foucault and R.D. Laing.

[3] For example, in 19th century America, drapetomania was the 'condition' that caused black slaves to run away from their masters, and in the USSR, political dissent was labeled a 'mental illness' (Szasz, 1961).

[4] Mourning, it has been argued, can now be diagnosed as major depressive disorder in the DSM-5, since it no longer rules out depression in response to the loss of a loved one (Horwitz and Wakefield, 2007).

## —1.1 Thomas Szasz's Conceptual Analysis

The concept of physical illness, at least for Szasz, necessarily implies it having a physical cause (1961). A physical cause entails that it occurs 'passively'. Indeed, it seems that one thing that all physical illnesses have in common is that they are things that 'happen to us', not things we 'do' that we have control over. No one 'chooses' to have cancer, it is something that just happens to us[5]. Szasz also held that the concept of physical illness implies that there is something wrong with the natural functioning of our bodies. One of the natural functions of the cell is to be integrated with the rest of our bodily functions. When cells grow uncontrollably, we have cancer. Such biological norms are given by nature, according to Szasz, as they exist independently of how we choose to evaluate them. How we evaluate cancer is up to us, but it is counter-intuitive that we dictate the natural functions of cells.

Szasz therefore claims that the concept of illness implies passivity and natural norms. But for Szasz, these features do not apply to the things we call 'mental illnesses'. Rather, mental illnesses are things that do not just happen to us like cancer: they are things where *what we do* forms a central part of the condition's onset, development, and/or treatment. What exactly Szasz thought that we do or do not have choice over in mental illness is not always clear, and whether we have free will, as he thought, is disputed and often dismissed as unscientific. But consider the fact that we would consider it ineffective to use psychotherapy to *treat* dementia. The reason is because we assume that dementia is not something that can treated through acceptance, changing maladaptive ways of thinking, or re-evaluating our circumstances, while we do think this is appropriate for conditions like depression following the loss of one's child. The implicit assumption here is that physical illnesses like dementia are not subject to change with the use of our minds over time, whereas mental illnesses like depression in a certain sense are. For Szasz, the concept of mental illness denies that we retain agency over how we use our minds[6].

Because mental illnesses involve our role as active agents and the choices we make, for Szasz mental illnesses are subject to ethical norms on how we ought to act. But for Szasz, ethical norms are not given by nature. Rather, ethical norms are provided by one's society. Such norms are much like the rules of a game, and playing games well for Szasz means playing the social roles that society provides (such as mother, ruler, teacher, etc…) well, and living up to our moral duties. However, playing a socially valued role for Szasz is only a necessary but not a sufficient condition for what we call mental health. For Szasz it is sufficient for mental health that the social role one plays is self-selected. But this also requires that one develop certain 'meta-values', such as autonomy and self-knowledge, in order to pursue the particular roles that one individually chooses to value.

---

[5] For Szasz, the passivity of illness also implies that medical interventions, such as surgery and medication, are appropriate forms of treatment, since illnesses are not subject to one's direct will to change long-term.

[6] For this reason then, medication and neurosurgery are inappropriate treatments for mental illnesses for Szasz.

Because according to Szasz rules of ethics are provided by society, his distinction between physical and mental illnesses seems to be between biological versus social norms. But while Szasz did hold this view too, this is not the essential distinction. What is essential is that the social norms for how *we* as free agents ethically ought to act are 'subjective', whereas natural norms for how the *body* ought to be are based on objective facts about our biology (or even natural facts about our own psychology). What exactly is meant by objective vs. subjective norms is not always so clear in this debate. Objective norms are often taken to mean universal norms for how we ought to be or live, denying that there are different social or individual values for how we ought to live, which is what many mean by subjective norms. Indeed, what is a good life for me might mean choosing a certain career, whereas a different career might make you happy. And there is no objective basis to say whether one career choice is better than the other if each make us individually happy. But no one really denies that there are such individual differences. Indeed, it seems to be an empirical objective fact that a certain kind of career will make you happy whereas the same career would not make me happy. But there is an underlying assumption here.

The assumption is that a life of happiness and a life free of needless suffering is indeed a good life, that this is what we ought to act for. But on what basis do we justify this assumption? Why ought we to act for our own happiness as opposed to living a life of suffering, perhaps because we think this will please God or serve society? If the claim that we ought to strive for some kind of life is some fundamental normative claim, what non-moral or non-circular claim can we appeal to justify living a certain way? Perhaps we cannot say why one ought to strive for well-being or suffering any more than chocolate instead of vanilla flavoured ice cream.

The problem here is David Hume's fact/value distinction, which holds that we cannot derive any normative claims for how we ought to live or act from objective facts without such normative claims already being assumed by our factual claims. For example, we may argue that, because animals suffer (factual claim) we ought not to eat them (normative claim). But the normative claim only follows because we are already assumed that suffering is bad in our factual claim. This is begging the question. A common conclusion that is drawn from this is that ethics is ultimately arbitrary, just a matter of preference. Indeed, if ethical norms are playing a game well by following the rules, then it follows that ethics is arbitrary since the rules of a game are simply agreed upon and therefore also arbitrary.

The view that ethics is arbitrary was widely held at this time and has also been attributed to Szasz. This is why Szasz rejects the concept of mental illness. It already assumes that, since there are certain facts about our biology and psychology that explain impairments and suffering, that these are objectively bad ways to live life. But since we cannot infer normative claims from such facts, mental health sciences and practices falsely assume that how we ought to live is based on objective norms. It is therefore irrelevant whether the suffering is normal, such as grieving, or severe, like schizophrenia. Since mental illnesses involves us as agents, mental health science and practice is fundamentally ethical and therefore subjective. The claim that mental illnesses are real illnesses denies this, it implies there are objective ethical norms, which for Szasz is a myth.

## —1.2 Normativism

In response to Szasz, some have argued that physical illnesses are just as much based on subjective values as mental illnesses are, instead of physical illnesses being based on objective norms of how the body ought to function. Cancer, then, would not be an objective illness. Nothing would be inherently bad about autonomous and destructive cells: cancer is just a certain state of the body that we don't like subjectively. This is simply pushing the fact/value distinction to its logical conclusion to also include biological norms over and above just ethical norms. Those who hold that illnesses in general are based on non-objective and therefore subjective norms are called Normativists (Sedgwick, 1973; Engelhardt, 1976; Goosen, 1980; Fulford, 1989).

But even if we allow for objective natural norms, like increased survival and reproduction, and subsequently attempt to make the factual claim that medical illnesses are reductions in biological fitness, then we cannot infer from this that we ought to value biological fitness. This would violate the fact/value distinction, since the inference already assumes that we ought to value biological fitness and that reductions in biological fitness are objectively bad for us. But this is begging the question. It suggests that reductions in biological fitness simply are not what we mean by medical illness. Sickle cell disease promotes overall biological fitness by offering increased immunity to malaria, but we still see it as a medical condition. We seek medical care not to increase our biological fitness for the promulgation of the species. Rather, we seek medical care to extend our lives, have children, and relieve ourselves of needless suffering, because we tend to value these things in life. As Goosen points out in defense of this position: "Medicine serves not the species nor the individual's lineage, but the individual patient" (1980, p. 113).

But if cancer is an illness only because we happen to value living longer lives free of suffering, then illnesses are just anything about our bodies that we do not like. But this cannot be what we mean by illness either, since there are other things we may not like about ourselves or our bodies that we do not see as illnesses, like being short or bald. Some Normativists, like Szasz, argue that illnesses are conditions that impair certain meta-values, such as autonomous, rational ordinary action, that allow us to pursue our subjective values. But now we face a problem. If we define autonomy, rationality, and ordinary actions too loosely, then we cannot rule out shortness because someone might see it as impeding ordinary action that they value. However, if we argue that in this case it is irrational, in the sense that that it does not impede autonomy, then we are being too specific. One would be implying that there are things we should and should not objectively value; it is not subjective after all. So while meta-values seem to be important to what we think illnesses are, they do not seem to explain what we really mean by illnesses because they fail to explain the distinctions we make between different subjective values. Another route would be to draw on certain facts about us. Indeed, it can't be a coincidence that we tend to value things that also overlap with reductions in biological fitness. We therefore, in some sense, need to figure out what this relationship is. Finding such facts does not violate the fact/value distinction, since they merely tell us what distinguishes health values from other things that we value.

## —1.3 Naturalism

Conceptual analyses have shown that illness involves subjective values, but that they are not sufficient to explain what we mean by illness. Those who claim that it is certain facts about us that make the meaning of illness distinct from other things we don't value are called Naturalists.

The most famous Naturalist account is Christopher Boorse's Bio-Statistical Theory (1975). While Naturalists like Boorse do not deny that illnesses are the subject of our negative evaluations, they claim that the concept of illness has a basis in objective facts about us that are completely independent of our evaluations of them. Illness has this basis in the concept of disease, a theoretical concept separated from medical practice. His official definition of disease is quite technical[7], but in essence for Boorse a disease is a reduction in present rates of biological fitness. However, as we saw, a reduction in biological fitness cannot be sufficient for something being an illness since this would violate the fact/value distinction. But for Boorse, what is also required for something to be an illness is that it is a disease that is not valued. This may appear to violate the fact/value distinction, but it is not intended to, since it is not claimed that we ought to value biological fitness as such. Rather, a disease being a necessary condition for illness only explains what kind of values illness refers to, namely, biological fitness rather than aesthetic values like eye colour. We only ought to value biological fitness if the consequences of a reduction in biological fitness, such as cancer, are something that we do not like. Illnesses are those reductions in biological fitness that we do not value, and this explains what we mean by illness.

The two main criticisms of Boorse's account are that there are counter-examples to his conceptual analysis and that his analysis of disease turns out to violate the fact/value distinction. While understanding diseases as reductions in biological fitness rules out baldness and other things we don't like about ourselves as distinct from the category of illnesses, this definition is now too narrow and rules out too much, such as sickle cell disease (which increases overall biological fitness). And even if all diseases involved a reduction in biological fitness, this already seems to assume that only the conditions that reduce our biological fitness ought to be valued. But this, again, seems to assume that there are objectively better ways to live, namely, that we ought not to like those conditions that reduce biological fitness. But this cannot be assumed, since it begs the question whether that is what we should do. Medical science and practice does after all then seem to assume fundamentally ethical norms. And since virtually all participants in this debate accept that ethical norms are subjective, then medical science and practice also are inherently subjective. Disease, as objective, would also be a 'myth' under such an understanding.

[7] Boorse defines disease as a failure to contribute to statically typical levels of biological fitness according to one's reference class.

## —1.4 Hybrid accounts

The third conceptual analysis in this debate tries to ground facts about us and their consequences on our values by combining Naturalism and Normativism into one concept, so-called Hybrid accounts. The most well-known account is Wakefield's Harmful Dysfunction Analysis (HDA) (1992).

Around the late 1970's, the concept of mental illness was being replaced with the concept of mental disorder. The concept of disorder is supposed to be equivalent to Boorse's concept of disease. But Wakefield does not distinguish between medical science as value-free and medical practice as value-laden. Medical science studies the conditions that medical practice treats, and so is equally determined by our social values. However, Wakefield argues that the concept of disorder also has some basis in objective facts about us. He grounds the concept of disorder in dysfunctions, but he has a different analysis of function than Boorse does. Rather than being failures to contribute to present levels of biological fitness, dysfunctions are failures of a mechanism to perform the function they were selected for in our evolutionary past. Consequently, this notion of dysfunction is only an explanation for *why* we think a mechanism is 'broken'. We still need a consequentialist concept in our analysis if we are to make dysfunctions relevant to what we value in medical practice. This is what Wakefield's concept of harm does. A disorder is a dysfunction that causes harm to someone, a harmful dysfunction. Indeed, we consider conditions to be medical illnesses because of their potential to cause ourselves harm. In turn, the concept of disorder constrains the relevant harms to those that are caused by dysfunctions, distinguishing disorders from other harms, like poverty or being not physically appealing. And since there are mental dysfunctions, those that cause harm are mental disorders. So mental disorders can exist, contra Szasz, and are grounded in both facts and values. However, values themselves are still not derived from such facts, according toe the fact/value distinction, only constrained by them.

The HDA does get closer to grounding the concept of disorder. It is harmful consequences of the breakdown of our bodies and minds, not their failure to contribute to fitness, that makes something a disorder. Indeed, a failure to function as designed by our evolutionary past does cause harm, such as the hearts' failure to pump blood in cardiac arrest. However, the HDA faces several problems. First, it is empirically doubtful that all mental disorders are also evolutionary dysfunctions. But even if they happened to be such dysfunctions, this cannot be what makes them medical disorders. If conditions like phobias or depression happened to be selected for, we would still see them as disorders because of their harm. It remains unclear why only harms that are caused by evolutionary dysfunctions would make something a disorder, but not other harms, like environmental mismatches. The deeper reason why traditional conceptual analyses have not been successful in grounding the concept of disorder is because of the following. If we maintain the fact/value distinction, which implies in this context that part of mental health norms are purely dependent upon the subject, that is, subjective, then no necessary conditions can ground the concept of disorder in objective facts. Values' contingent relationships to facts cannot be made necessary by stipulation. We cannot try to bridge the fact/value distinction while simultaneously trying to keep facts and values apart.

## —2.0 Future directions

The debate over the concept of mental illness has been going on since the 1960s, and has assumed the validity of traditional conceptual analysis and the fact/value distinction. But these assumptions have been challenged by contemporary philosophy. Despite this, the concept of mental illness debate has remained relatively isolated from these developments. In the second half of this paper I will show how bringing contemporary philosophy to bear on this debate has promise to move it forward and make it more relevant to mental health science and practice.

### —2.1 Traditional conceptual analysis

Analysing how we use a concept can tell us what we mean by a concept, such as analysing how our concept of illness is used to refer to deviations from some kind of norm regarding the health of human beings. But the way that conceptual analysis is traditionally used in the 'concept of illness debate' assumes that analysing the usage of a concept is often also sufficient to tell us about how we should use such concepts, such as whether or not we should apply the concept of illness to the mind. This is a kind of conceptual conservatism about the nature of medical concepts, namely, that concepts already have 'within them' the way in which such concepts should be used and that we just need to clarify such usage to determine their correct application. This is reflected in how intuitions are used both as evidence that a conceptual analysis correctly describes how we use the concept of illness and therefore prescribes how we should use it. But one of the problems of conceptual conservatism and appealing to intuitions is that there does not seem to be a systematic means to settle which intuitions are worth preserving when they conflict. For example, Szasz takes it as intuitive that illnesses are passive in nature and therefore it is highly counterintuitive to apply the concept of illness to learned behaviours. Yet Szasz's critics do not take it as counter intuitive to apply the concept of illness to learned behaviours. Naturalists take it as counterintuitive if illnesses refer merely to values independent of facts and normativists take it as counterintuitive if illnesses refer merely to facts independently of values. The success of hybrid accounts is in large part due to their ability to 'balance' more of these intuitions. But there is a deeper problem than simply how we balance our intuitions when they conflict. What if our intuitions are not a reliable source for describing the usage of concepts, nor for how we should use concepts? This is one of the conclusions that has come out of recent developments in the study of philosophical methodology in academic philosophy. Some argue that we should still use intuitions as a source for describing and prescribing how we use concepts, but that we should do empirical research on those intuitions because they may differ by culture (Machery, 2017). Others have questioned the use of intuitions and the conceptual conservatism it implies as such, arguing that we should be more concerned with developing our concepts to fulfil certain ends rather than preserving and balancing our intuitions for their own sake (Eklund, 2015). We may want to conceive of old age as a disease because many diseases are a consequence of ageing. But assumptions about methodology are not the only thing being challenged in this debate.

## —2.2 Facts and values

What exactly the fact/value distinction means in the context of the concept of illness debate is not always clear. In the particular case of the concept of mental illness, it seems to mean that, when it comes to ethical and health norms, we cannot appeal to objective facts alone to say which are better or worse. Such norms are matter of social custom or personal preference, it is argued, which is what we mean by 'subjective' rather than 'objective'. But if these terms are equally unclear, their purported acceptance has legitimate motivations that need to be acknowledged.

First, there *are* differences in values between both social groups and individuals in how to live a good life where we may indeed not be able to say which are better or worse, such as individual values concerning occupational choices and different artistic expressions in cultures. But the idea that there are objective norms is historically often taken to imply that there is one and only one way to live a good life, such as the life of the philosopher according to Aristotle. To the extent that we have been wrong in the past about what constitutes a good life, like being Christian and straight, we may therefore want to reject any postulation of objective ethical and mental health norms. Second, claims about objective norms have historically been used as the basis to oppress individuals or groups of people that deviate from those norms. We therefore have legitimate political concerns about claims of objective norms.

But such a political concern is distinct from the actual existence of objective and subjective norms, and subjectivism about norms is not the only way to deal with such concerns. As for the fact/value distinction itself, when we ask for concrete examples and attempt to push it to its logical conclusion, problems begin to emerge in our conceptual analysis of mental illness. This is because it is simply not true that all that we mean by mental illnesses are just deviations from our social agreements and personal preferences in the sense that disease is equivalent to other social violations, like failures of etiquette. Even *if* subjective norms ground the concept of mental illness, they fail as a conceptual tool for distinguishing between mental health norms and other subjective norms, such as other negative evaluations of mental traits and dispositions that we do not like, like being really annoying.

The solution that is often sought for this 'run-away' subjectivism is to add constraints as a necessary basis for mental health and illness. These can be 'meta-values' that make the pursuit of specific values possible, or objective norms such as evolutionary norms. But to the extent that such constraints are supposed to limit such cases, they seem incompatible with the fact/value distinction. This is because they are based on certain facts about us as rational and social animals with various requirements in order to function well based on how we use such capacities, such as acquiring knowledge of ourselves and gaining autonomy. That there are facts about us that we share with other animals and most other human beings, as well as unique individual differences relevant to mental health, is clear. The question now is what those shared facts and individual differences are. But at this point, what it means exactly for norms to be objective or subjective is less clear than before. We need to clarify what we mean by *these* concepts if we are to move this debate forward rather than assuming that we know what we are talking about.

There is currently a debate in ethics that deals specifically with just this question of how objective vs. subjective factors constitute our psychological well-being (Tiberius, 2010). This debate is part of a larger development in ethics since the 1970's, which concerns the revival of virtue ethics in the Ancient Greek philosophical tradition (Anscombe, 1958; Foot, 2001). These developments also call into question the fact/value distinction by arguing that what we ought to do to live good lives can be determined by what human beings require to flourish based on objective facts about human nature, such as our biology and psychology. Unlike previous objective accounts of the good life, these accounts of well-being attempt to accommodate a higher diversity of ways of living good lives (Nussbaum, 2007), while still adding certain objective standards, such as dignity, much like the Normativists' meta-values in the mental illness debate.

This debate has also led to research on the relationships between virtues, such as self-control, resilience, and their relationship to mental health. The fruits of this work are new frameworks for therapeutic interventions that can help empower and bolster the agency of patients, making them more active in their own care (Pearce and Pickard, 2010). This requires us to fundamentally rethink the distinctions that underlie psychiatry and the distinctions those fighting mental illness stigma make, such as the distinction that underlies the idea that mental illness is 'chemical, not character' and cannot be both. Thomas Szasz may therefore have been right that some mental illnesses have a moral component, but wrong to argue that this makes them unable to also be considered as medical disorders (with biological components).

An important feature of these developments is its interaction with the empirical sciences, such as psychological work on theories of well-being, character-development, and self-control (Pickard, 2013). Such work, however, still requires some form of conceptual analysis. Analysis can help us think more clearly about the underlying conceptual assumptions that determine how we interpret empirical data and relate it to our larger conceptual scheme, as opposed to accepting or leaving unclear how we think of objective and subjective norms. But as we have seen, traditional conceptual analysis tends to assume that most of our concepts already tend to map onto clear distinctions in reality, and that we should preserve our intuitions at the potential cost of developing more useful concepts or to develop the concepts that we do have in more useful ways. To the extent that this method has been called into question by philosophy, it is no longer adequate if we want an empirically informed debate about mental health and illness. In other words, we should be more willing to sacrifice certain intuitions about the distinction between different problems-in-living if doing so means that doing so allows us to rethink them.

In response to the criticisms of traditional conceptual analysis, a new debate is going on about alternative philosophical methods that are more empirically informed and practically oriented, such as philosophical explication in the tradition of Rudolf Carnap, which focuses not primarily on what we mean by concepts, but also how to improve the role that they play in science. Another example is conceptual engineering, which seeks to address the broader question of what constitutes having a good concept in various fields like philosophy, policy, and science (Eklund, 2015). Drawing on such research can help us develop better standards for the clear thinking that mental health sciences and practices need, such as developing our understanding of the relationship between different problems of living in medicine and beyond.

**Conclusion**

The concept of mental illness debate was a response to legitimate concerns regarding abuses of psychiatry and the medicalisation of normal suffering and differences. Traditional conceptual analysis assumes that our underlying conceptual scheme that our common sense intuitions and existing medical expertise are based on can tell us what mental illness is. However, whether they *can* is precisely what is at issue: it cannot be assumed and needs to be argued for. In particular, our assumption of the fact/value distinction can no longer be taken for granted given contemporary debates in ethics. And if the issue is about general norms vs. individual differences in mental health, then this is something that science can objectively study. But to the extent that this requires us to rethink our assumptions, traditional conceptual analysis is problematic since it assumes that our underlying conceptual scheme only needs to be clarified as opposed to being challenged and developed. Drawing on current philosophical methods that are more empirical and practically oriented will therefore help make the concept of mental health debate relevant to mental health science and practice.

**References**


Anscombe, 1958. Modern Moral Philosophy. *Philosophy* 33:124, p. 1-19.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44, p. 542–573.

Medlin, B. 1957. Ultimate principles and ethical egoism. *Australasian Journal of Philosophy.* 35:2, p. 111-118.

Eklund M. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In: Daly C. (eds.) The Palgrave Handbook of Philosophical Methods. Palgrave Macmillan, London.

Engelhardt, H. T. 1976. Ideology and Etiology. *The Journal of Medicine and Philosophy* 1:3, p. 256-268.

Foot, P. 2001. Natural Goodness. Oxford University Press, Oxford.

Fulford, K. W. M. 1989. Moral Theory and Medical Practice. CUP, Cambridge.

Goosens, W. K. 1980. Values, Health, and Medicine. *Philosophy of Science* 47:1, p. 100-115.

Horwitz, A. V., and Wakefield, J. C. 2007. The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder. Oxford University Press, Oxford.

Machery, E. 2017. Philosophy within its proper bounds. Oxford University Press, Oxford.

Murphy, D. 2006. *Psychiatry in the scientific image*. Cambridge, MA: MIT Press.

Nussbaum, M. C. 2007 Frontiers of Justice: Disability, Nationality, Species Membership: The Tanner Lectures on Human Values. Belknap Press: An Imprint of Harvard University Press.

Sedgwick, P. 1973. Illness: Mental and Otherwise. The Hastings Center Studies, 1:3, p. 19-40.

Szasz, T. S. 1960. The Myth of Mental Illness. The American Psychologist 15:2, p. 113-118.

Szasz, T. S. 1961. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins.

Szasz, T. S. 2010. Psychiatry, Anti-Psychiatry, Critical Psychiatry: What Do These Terms Mean? *Philosophy, Psychiatry, & Psychology* 17:3, p. 229-232.

Tiberius, V. 2010. Well-Being. In The Moral Psychology Handbook. Edited by John M. Doris and The Moral Psychology Research Group. Published to Oxford Scholarship Online.

Pearce, S. and Pickard, H. 2010. Finding the will to recover: philosophical perspectives on agency and the sick role *Journal of Medical Ethics* published online.

Pickard, H. 2013. Psychopathology and the ability to do otherwise. *Philosophy and Phenomenological Research* p. 1-29.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundaries between biological facts and social values. *American Psychologist* 47:3, p. 373–388.

Zupan, M. L. 1973. Is mental illness a myth? *Reason* 5/61:3, p. 4-11.

**Szasz and the Disease Versus Choice Distinction: an Unanswered Question for Medicine**

Thomas Szasz infamously argued that mental illnesses are not real[8] illnesses[9] and that the concept of mental illness functions to justify the social control of deviant behaviour (Szasz, 1960 and 1961a). Mental illnesses are not real illnesses, according to Szasz, because of the meaning of the concept of illness as bodily illness (Szasz, 2011). Szasz's conceptual analysis is often read as a claim that real illnesses are biological and value-free while mental illnesses are not. On such a reading, Szasz was primarily arguing against the biomedical model of mental illness. The two practices that Szasz argued against most as being justified by the concept of mental illness were involuntary treatment and the insanity defence (Szasz, 2004). On such a reading, Szasz argued primarily against psychiatry as a tool for social control and as an illegitimate branch of medicine.

Critics have responded to Szasz by arguing that the concept of mental illness does not play the role that Szasz thinks it does in practices like involuntary treatment (Pies, 1979), and that concerns about social control can be addressed by defining mental illness as (in part) harmful to the individual, rather than to society at large (Spitzer et al., 1978). Critics have also argued against Szasz that illnesses are not defined as bodily illnesses and have proposed alternative definitions of illness that are not exclusively biological nor entirely value-free (Kendell, 1975; Boorse, 1975; Pies, 1979; Fulford, 1989; Wakefield 1992). Most critics of Szasz claim that mental illnesses are therefore real illnesses and, thus, that psychiatry is a legitimate branch of modern medicine.

I argue that these readings fail to grasp Szasz's primary argument against the concept of mental illness, which was that real illnesses are involuntary while mental illnesses are defined as voluntary, and that the concept of mental illness is thereby a rejection of the disease versus choice distinction. I argue that Szasz's primary criticism of the concept of mental illness, and thereby psychiatry, is that it rejects the disease versus choice distinction. While I argue that Szasz simply took for granted that medicine ought to maintain a disease versus choice distinction, I also argue that whether medicine ought to maintain a disease versus choice distinction remains an unanswered question, as evidenced by debates about the status of addiction (Heyman, 2009).

This paper is divided into four parts. In part one, I present Szasz's analysis of the function of the concept of mental illness as a denial of agency and how it relates to his claim that mental illnesses are not real illnesses. In part two, I argue that Szasz's analysis of illness as biological and value-free cannot explain why Szasz is also critical of conceptions of mental illness that do not claim to be biological or free of values. In part three, I argue that what explains Szasz's rejection of the concept of mental illness is that he regards them as voluntary, and that he regards real illnesses as involuntary. I argue that Szasz holds that medicine ought to maintain a disease versus choice distinction. In part four, I criticise Szasz for taking the disease versus choice distinction for granted, and argue that whether or not medicine ought to maintain a disease versus choice distinction remains an unanswered question, as evidenced by the debates about addiction.

---

[8] Szasz's used the claim that 'mental illness is a myth' and that 'mental illnesses are not real illnesses' interchangeably (Szasz, 1961).

[9] Szasz used the concept of illness interchangeably with disease and disorder. So while others in the literature distinguish between such concepts (e.g., Boorse, 1975 and Wakefield, 1992), I too will use such concepts interchangeably for the purposes of this paper.

## —1. Thomas Szasz's functional analysis of the concept of mental illness

Szasz's position on mental illness consists of two central claims: (1) that mental illnesses are not real illnesses, and (2) that the concept of mental illness functions to deprive people of their agency (Szasz, 1961a and 2010). These two claims are based on Szasz's analysis of the meaning *and* function of the concept of mental and bodily illness respectively[10]. Here, in part one, I present Szasz's functional analysis of the concept of mental illness and argue that its *primary* function for Szasz is the denial of agency, with social control only as a derivative function. This primary function motivates Szasz's argument that mental illnesses are not real illnesses. I will argue, however, that what exactly it means for the concept of mental illness to 'function to deny agency' is unclear. I will argue that it is also unclear how this claim is related to his other claim that mental illnesses are not real illnesses. I will take up these two issues in parts two and three.

Szasz stated that, if it were not for the function of the concept of mental illness, he would not have bothered arguing that mental illnesses are not real illnesses[11]. But Szasz attributed many functions to the concept of mental illness, so it is essential to determine what these are, how they relate to each other, and how they relate to his claim that mental illnesses are not real illnesses.

The function of the concept of mental illness that most people remember Szasz for focusing on is its function in justifying various forms of social control and 'excuses', especially in the form of involuntary treatment[12] and the insanity defence[13]. Critics, however, have pointed out that the concept of mental illness does not in fact play this role[14]. But Szasz's analysis of the social control function of the concept of mental illness was much wider than this. He also argued that the concept of mental illness exerted social control even when it does not lead to involuntary treatment. This is because the concept of mental illness pathologises some behaviours as abnormal (Szasz, 1961a). Critics have responded that these cases can be avoided by defining mental illness in part as harmful to the individual, not to society as such (Spitzer et al., 1978).

At the same time, Szasz also attributed functions to the concept of mental illness *other* than social control. For example, Szasz argued that some people label *themselves* as mentally ill without being ill in order to enter into the sick role and access medical resources, so-called malingering (Szasz, 1961a). Szasz also argued that the concept of mental illness functions to bring medical resources to legitimate non-medical problems that society would otherwise ignore, such as social disadvantage. Szasz thought the proper solution to this would be to expand our scope of care to the non-ill rather than endlessly expand the concept of illness (Szasz, 1961b).

---

[10] Szasz himself refers to this distinction as the informational/cognitive vs strategic meaning of the concept of mental illness (1961b).

[11] "If the legal and social consequences of being diagnosed "mentally ill and dangerous to self or others" were as theoretically unimportant and socially inconsequential as diagnosing him as, say, "having a mild fungus infection of his toenails," then I would not care whom psychiatrists label mentally ill" (Szasz, 2004a p. 356.)

[12] "…I maintain that the principal meaning and use of such terms [as mental illness] is [sic] strategic, justifying psychiatric coercions and excuses, epitomized by involuntary hospitalization and the insanity defense" (Szasz, 2004b, p. 132).

[13] "All tests of criminal responsibility rest on the premise that people 'have' conditions called 'mental diseases' which 'cause' them to commit criminal acts. The value of these tests thus hinges on the soundness of this underlying concept" (Szasz, 1967 p. 272-273).

[14] These practices are justified by determining whether an individual is legally component (see Pies, 1979 and 2004 and 2019).

But most of these functions that Szasz attributed to the concept of mental illness can be traced back to a *primary* function that Szasz was most concerned about, namely, that the concept of mental illness functions to undermine or deny agency[15]. Indeed, most of the functions that Szasz attributes to the concept of mental illnesses, such as social control, involuntary treatment, the insanity defence, entering the sick role, are different manifestation of undermining or denying agency[16]. For Szasz, then, to say that someone has a mental illness is simply to say that they lack agency or have diminished agency. But it isn't exactly clear what Szasz means by the concept of mental illness functioning to deny agency. At times, he simply seems to say that the concept functions to deny agency only in the person who is labelled as mentally ill. But at other times he seems to claim that the concept of mental illness functions to undermine or deny *belief in agency as such*, irrespective of whom the concept is applied to[17]. The important follow-up question in this regard is how exactly Szasz's functional analysis is related to his claim that mental illnesses are not real illnesses. Scholars have proposed different ways of making sense of how these two claims are related. Here are some of the ways that Szasz's position has been read.

One possible way to read Szasz here is that he does not fundamentally object to the function of the concept of mental illness *as such*, but only objects to this function in the case of mental illnesses in particular because mental illnesses *happen* to not be real illnesses. However, Szasz himself rejected this reading of his position. Szasz argued that people who are bodily ill should not be coerced in the way that he thinks the mentally ill are[18].

According to Szasz, the concept of bodily illness also functions to deny agency, but only in a very delimited way over very specific capacities. In contradistinction, the concept of mental illness, according to Szasz, is a *global* denial of agency in the person labelled as mentally ill:

"In general, then, while we do not consider medical patients to be responsible for being ill, we do consider them, despite their illness, to be responsible for what they do with their lives. This is especially so when the illness is chronic, in which case we typically consider the patient responsible for managing his disease [….] In contrast, mental illness typically confers precisely this sort of total nonresponsibility on its victims. Why should this be so? Why do psychiatrists and the law not treat psychotics like physicians and the law treat diabetics — regarding them as not responsible for their disease, but responsible for their deeds? The fact that they do not reveal what, *inter alia*, the idea of mental illness really means: namely, nonresponsibility — not only for one's condition, but for virtually any aspect of one's conduct as well" (1993, p. 268)

---

15 "Psychiatrists understand that their entire enterprise hinges on society's acceptance of the proposition that human beings diagnosed as mentally ill have a brain disease that deprives them of free will" (1991, p. 1576).

16 "The dependence of moral agency on mindedness renders the judgment of impaired mindedness–that is, the diagnosis of "mental illness"–of paramount legal and social significance" (Szasz 2001, p. 300).

17 "…the concept of mental illness also undermines the principle of personal responsibility, the ground on which all free political institutions rest" (Szasz, 1997 p. 78).

18 "I clearly state that since somatic pathology per se is not a sufficient condition for depriving a bodily ill person of moral agency, it should not be a sufficient condition in the case of a mentally ill person either (Szasz, 1987, p. 160)

Szasz does not argue that mental illnesses are not real illnesses because he thinks that the particularistic-agency-denying function of the concept of bodily illness is being inappropriately applied to things that are not real illnesses in the case of mental illnesses. This is because Szasz thinks that the concept of mental illness functions quite distinctly from the concept of bodily illness in denying agency globally. It is this unique 'global' function that Szasz is objecting to.

Therefore, another way to read Szasz is that his argument that mental illnesses are not real illnesses is simply just an argument against this kind of function. This is what Engelhardt calls the 'methodological interpretation' of Szasz's position (Engelhardt, 2004). Engelhardt argues that what is central to Szasz's position is that there is a fundamentally different methodological approach to treating individuals as bodies as opposed to agents, and that because agency plays some role in many mental illnesses, we should not treat them as literally ill, that is, as non-agents, as we do with bodily illnesses[19]. Engelhardt argues that Szasz's claim that mental illnesses are not real illnesses is therefore not a substantive metaphysical or empirical claim at all about what mental illnesses are, but simply a call to treat people that we call mentally ill as agents as opposed to bodies. Engelhardt, however, argues that one can maintain that mental illnesses are real illnesses while at the same time maintaining that there is a methodological distinction between medicine and addressing problems that require agency such as in psychotherapy (Engelhardt, 2004).

Is Engelhardt right about his interpretation of Szasz? Szasz's own response to Engelhardt is that it is 'partly right' that his claims about mental illness are neither metaphysical nor empirical[20]. That is, *Szasz holds that he is not fundamentally trying to make a claim about the nature of the things we call mental illnesses*. However, Szasz rejects the notion that he is not making any kind of substantial claim whatsoever. Rather, Szasz argues that his claim that mental illnesses are not real illnesses is fundamentally an analytical claim about the meaning of the concepts of bodily and mental illness respectively. In particular, Szasz claims that his position that mental illnesses are not *and cannot* be real illnesses is derived from his analysis of the meaning of the concept of illness as bodily illness. That is, because of what both the concepts of bodily illness and mental illness *mean*, mental illnesses literally cannot be real illnesses[21].

This raises two questions. First, which meaning(s) of the concept of illness and mental illness does Szasz have in mind when he argues that mental illnesses are not real illnesses? Second, how exactly does this analysis of their meanings relate to and clarify the meaning of Szasz's claim that the concept of mental illness functions to deny or undermine agency? These two questions will guide my analysis of Szasz's conceptual analysis.

---

[19] "Szasz's core claims are best viewed as methodological and moral, not empirical or metaphysical. Szasz should be read as holding that, for important moral and cultural reasons, a methodological separation should be established between the role of psychotherapists and the role of physicians" (Engelhardt 2004, p. 366).

[20] "[Engelhardt] states that my claim that mental illness is a myth "should not be regarded as an empirical or metaphysical claim that mental diseases do not or cannot exist." This is partly right. My claim asserts an analytic, not a synthetic, truth; as such, it is not based on empirical observation" (Szasz, 2004, p. 376).

[21] "In *The Myth of Mental Illness*, I argued that mental illness does not exist not because no one has yet found such a disease, but because *no one can find such a disease*: the only kind of disease medical researchers can find is literal, bodily disease" (Szasz 2004, p. 322). And; "When, in 1960, I first asserted that metal illness is a myth, I meant to remind people that, according to strict medical definition, disease is a predicate of *(human) bodies*. If we grant that definition, then we need not examine any particular person to know that he does not have a mental illness. The mind can be ill only in a metaphorical sense" (Szasz 2004, p. 377).

## —2. Thomas Szasz's conceptual analysis: real illnesses as biological and value-free

Szasz claimed that mental illnesses are not real illnesses because the meaning of the concept of mental illness is incompatible with the meaning of the concept of bodily illness. However, Szasz attributed various meanings to the concept of bodily illness, namely, that bodily illnesses are (1) biological, (2) value-free, and (3) involuntary. Here I argue that Szasz's claim that mental illnesses are not real illnesses is not primarily based on real illnesses being biological or value-free.

*Diseases as 'biological' or having 'lesions'*

Szasz argued mental illnesses are not real illnesses because of the meaning and definition of the concept of bodily illness. One feature of bodily illnesses that Szasz stressed as differentiating them from mental illnesses is that bodily illnesses have lesions (Szasz, 1961). By lesions Szasz meant anatomical or physiological alternations in the human body that are *in principle* observable and objectively discernible[22]. On this reading of Szasz, he would therefore regard the claim that mental illnesses are real illnesses as an empirical claim that mental illnesses *do in fact have lesions*. Claims of lesions could explain the function of the concept of mental illness in denying agency since Szasz also claims lesions could explain behaviour as involuntarily.

On this reading of Szasz's argument, the claim that mental illnesses are not real illnesses is therefore the claim that *they lack lesions*. So if mental illnesses do not have lesions, this reveals for Szasz that the concept of mental illnesses only functions to deny agency in others in order to justify using coercion against them. On the basis of this reading of Szasz many have criticised him by arguing that disease is not defined by lesions at all, and so his argument that mental illnesses are not real illnesses because they lack lesions fails (Kendell, 1975 and 2004; Pies, 1976, 2004, and 2019; Spitzer el al., 1978)[23]. This reading of Szasz, however, is mistaken for three reasons.

First of all, Szasz's view that bodily illnesses are defined as having lesions cannot explain why he thought that the concept of mental illness functions to deny agency, as he also attributed this function to non-biological conceptions of mental illness, such as certain conceptions in psychoanalysis[24]. His view that the concept of mental illness functions to deny agency therefore does not rest solely on his point about lesions.

---

[22] "Until the middle of the nineteenth century, and beyond, illness meant a bodily disorder whose typical manifestation was an alteration of bodily structure … [a] lesion, such as a misshapen extremity, ulcerated skin, or a fracture or wound" (Szasz, 2010, p. 11).

[23] "The various assertions that what psychiatrists regard as mental illnesses are nothing of the kind have all been based on the argument that no physical basis has ever been demonstrated in these conditions, and that some kind of lesion is essential to [p. 313] establish the presence of disease…. The argument of these writers are therefore all based, wittingly or unwittingly, on a concept of disease which has been abandoned not just by psychiatry but by medicine as a whole" (Kendell, 1975 p. 312-313).

[24] "Although differing in certain ways, old-fashioned asylum psychiatry, psychoanalysis, and modern biological psychiatry thus all agree on the all-important point, that the behavior of the mentally ill person is strictly *determined*: such a person has no free will and is therefore not responsible for his actions" (Szasz, 1987 p. 245).

Second, Szasz did not claim that mental illnesses are not real illnesses because the things they refer to *fail* to have lesions. Indeed, Szasz claimed that, if lesions are found for mental illnesses, this only showed that they are bodily illnesses, not that mental illnesses are real illnesses[25]. Rather, Szasz argued that bodily illnesses are *defined* as having lesions whereas he claimed that mental illnesses were instead defined as diseases by so-called functional criterion, such as suffering and disability[26]. Szasz, however, took such functional criteria to be too broad and vague to distinguish diseases from non-diseases. Thereby, such a criterion turns suffering of any kind into a potential disease. Szasz advocated for the lesion criterion because he thought it would prevent unnecessary medicalisation and limit the arbitrary power of medicine and institutional psychiatry[27]. Szasz's criticism of mental illness as 'not having lesions' was therefore not primarily a criticism that the things we refer to as mental illnesses do not have lesions, but a criticism that the criterion of illness was being changed to include any and all suffering.

Third, lesions and suffering are not mutually exclusive criteria and are in fact compatible: for example, some lesions *are* aspects of suffering/disability and vice-versa. Lesions therefore cannot be central to Szasz's argument that mental illnesses are not real illnesses: in essence Szasz claims that his argument is an analytical claim that the definition of bodily illness excludes mental illnesses from being real illnesses.

So while Szasz did argue that lesions are and ought to be part of the definition of illness, this was not ultimately what grounded his claim that mental illnesses are not real illnesses. Nor does this explain Szasz's functional analysis of the concept of mental illness. This is because Szasz also attributed the agency-denying function of the concept of mental illness to non-lesion conceptions of mental illness by psychoanalysts (see footnote 17). Moreover, his main argument was not that the things that the concept of mental illness refers to do not have lesions, but rather that mental illnesses are not *defined* as having lesions. Since lesions and functional criteria are compatible, we will have to look elsewhere for the basis of Szasz's two central claims.

---

[25] "Many scientists, doctors, and laypeople believe that if a genetic defect caused a mental illness or a lesion was found in the brains of mentally ill patients, it would prove that mental illnesses exist and are like any other disease; this is not so. If mental illnesses are diseases of the central nervous system, they are diseases of the brain, not the mind. If mental illnesses are the names of (mis)behaviours, they are behaviours, not diseases. A screwdriver may be a drink or an implement. No amount of research on orange-juice-and-vodka can establish that it is a hitherto unrecognised form of a carpenter's tool" (Szasz 1991, p. 1574).

[26] "It is important to understand clearly that modern psychiatry—and the identification of *new* psychiatric diseases—began not by identifying such diseases by means of the established methods of pathology, but by creating a new criterion of what constitutes disease: to the established criterion of detectable alteration of *bodily structure* was now added the fresh criterion of alteration of *bodily function;* and, as the former was detected by observing the patient's body, so the latter was detected by observing his behavior. This is how and why conversion hysteria became the prototype of this new class of diseases—appropriately named "mental" to distinguish them from those that are "organic," and appropriately called also "functional" in contrast to those that are "structural." Thus, whereas in modern medicine new diseases were *discovered,* in modern psychiatry they were *invented.* Paresis was *proved* to be a disease; hysteria was *declared* to be one" (Szasz, 1961 p. 12 original emphases).

[27] "The Virchowian standard is fixed by biological-physical criteria, limiting the medical system from arbitrarily expanding its scope and hence its power. Neither doctors, patients, politicians, nor any other interested parties can create diseases by manipulating the language" (Szasz, 2006 p. 333).

*Diseases as 'value-free'*

Another feature that Szasz argues distinguishes between bodily from mental illnesses is the kind of norms that they are deviations from. Szasz defines bodily illnesses as deviations from bodily norms that can be stated in anatomical and physiological terms[28]. Mental illnesses, Szasz argues, are defined in terms of deviations from psychological, social, legal, and moral norms[29].

Szasz was especially drawing on the first and second editions of the Diagnostic and Statistical Manual (DSM) of Mental Illnesses (APA, 1952 and 1968), which at that time was still very much a product of the dominant influence of psychoanalysis in American psychiatry. Psychoanalysis, at that time, as it still does today, does not describe mental illnesses in terms of deviations from biological norms, but describes mental illnesses in terms of deviations from psychological and social norms, such as psychic forces of the id, ego, and super-ego.

However, as we saw earlier, Szasz's criticism of mental illness did not just apply to psychoanalysis, but to all forms of psychiatry, including biological psychiatry. So how does Szasz's distinction between bodily and mental illnesses apply as a criticism of biological psychiatry, which presumably would define mental illnesses as deviations from biological norms?

Szasz does *seem* to make another distinction between bodily and mental illnesses. On the one hand, norms from which bodily illnesses are deviations, he claims are relatively more free from societal norms and values, i.e., are more value-free. On the other hand, mental illnesses are deviations from norms that are charged with societal values, i.e., are relatively more value-laden[30]. Many scholars have read Szasz as claiming that real illnesses are value-free whereas mental illnesses are value-laden (Boorse, 1975; Wakefield, 1992; Loughlin and Miles, 2015).

This reading of Szasz can provide an explanation of both his central claims. The function of the concept of mental illness as denying agency is explained by societal norms being 'passed off' as value-free diseases as a means for socially controlling others. And if bodily illnesses are defined as value-free, and mental illnesses as defined as value-laden, then mental illnesses cannot be real illnesses by definition, under the assumption that completely value-free phenomena are in a disparate category than those that are value-laden. This reading of Szasz could explain why he argued specifically against non-biological forms of psychiatry, if such forms also argued that the concept of mental illness is value-free. Indeed, I also used to read Szasz as primarily making a claim that illnesses are value free and mental illnesses are not (Schuur, 2019).

---

[28] "The concept of illness, whether bodily or mental, implies deviation from some clearly defined norm. In the case of physical illness, the norm is the structural and functional integrity of the human body. Thus, although the desirability of physical health, as such, is an ethical value, what health *is* can be stated in anatomical and physiological terms" (Szasz 1960, p. 114 original emphasis).

[29] "What is the norm deviation from which is regarded as mental illness? This question cannot be easily answered. But whatever this norm might be, we can be certain of only one thing: namely, that it is a norm that must be stated in terms of psycho-social, ethical, and legal concepts. For example, notions such as "excessive repression" or "acting out an unconscious impulse" illustrate the use of psychological concepts for judging (so-called) mental health and illness. The idea that chronic hostility, vengefulness, or divorce are indicative of mental illness would be illustrations of the use of ethical norms (that is, the desirability of love, kindness, and a stable marriage relationship)" (Szasz 1960, p. 114).

[30] "…whereas bodily disease refers to public, physicochemical occurrences, the notion of mental illness is used to codify relatively more private, sociopsychological happenings of which the observer (diagnostician) forms a part. In other words, the psychiatrist does not stand apart from what he observes, but is, in Harry Stack Sullivan's apt words, a "participant observer." This means that he is *committed* to some picture of what he considers reality — and to what he thinks society considers reality — and he observes and judges the patient's behavior in the light of these considerations" (Szasz 1960, p. 116-117 original emphasis).

On the basis of this reading, critics have argued that Szasz is wrong that there is a fundamental difference between the norms from which bodily and mental illnesses are deviations (Boorse, 1975; Fulford, 1989; Wakefield, 1992). However, there are several problems with this reading of Szasz. First, this reading cannot fully explain why Szasz thinks the concept of mental illness functions to deny agency *directly*. Rather, all this reading can explain is why 'passing off' value-laden societal norms as value-free norms can lead to indirect social control. Moreover, one could in principle agree with Szasz's critics that the norms from which bodily and mental illnesses are deviations are the same sort of norms, but still reject the concept of mental illness on the basis that it rejects agency. In other words, the issue of values and of agency here are conceptually independent. For example, one could hold that addiction is not a disease because it is involuntary, regardless of whether one regards the norms underlying diseases or moral problems as objective or subjective. Second, Szasz *himself* rejections this interpretation of his position. In particular, Szasz claims that *both* bodily and mental illness are value-laden, that is, determined by social norms. Their difference is the *kind* of value-laden norms they deviate from:

"…that the concept of disease contains an evaluative element is self-evident. . . . The crucial difference between lesion *qua* bodily disease and behavior *qua* mental disease is not that one is a value-free biological fact and the other a value-laden social construct. Both are value-laden social constructs. . . . The crucial difference between bodily disease and mental disease is that what counts as a somatic pathology is based on a judgment of how the *body ought to function*, whereas what counts as psychopathology is based on a judgment of how the *person ought to function*" (Szasz 2000, p. 9 original emphasis).

One major implication of mental illnesses being deviations from persons and their actions is that mental illnesses are more 'moral' in nature than bodily illnesses. But since 'moral' is often read as 'value-laden' or 'subjective', this has led many people (including myself) to misread Szasz as claiming that mental illnesses are not real illnesses because real illnesses are value-free and mental illnesses are value-laden. Rather, the main distinction seems to be between the *morally* value-laden norms of mental illnesses and the *non-morally* value-laden norms of bodily illnesses[31]. What Szasz means by this distinction I will turn to in part three. But what is clear at this point is that neither Szasz's claim that mental illnesses are not real illnesses nor his claim that the concept of mental illness functions to deny agency can be explained by his claims about real illnesses being value-free and mental illnesses being value-laden, since he never claimed this. I now turn to part three, where I argue that what explains Szasz's claims that mental illnesses are not real illnesses and that the concept of mental illness functions to deny agency is based on his view of the concept of illness as meaning 'involuntary' and mental illness as meaning 'voluntary.'

[31] "Human behavior is fundamentally moral behavior. Attempts to describe and alter such behavior without, at the same time, coining to grips with the issue of ethical values are therefore doomed to failure. Hence, so long as the moral dimensions of psychiatric theories and therapies remain hidden and inexplicit, their scientific worth will be seriously limited. In the theory of personal conduct which I have proposed—and in the theory of psychotherapy implicit in it—I have tried to correct this defect by articulating the moral dimensions of human behaviors occurring in psychiatric contexts" (Szasz, 1961 p. 263)

## —3. Thomas Szasz and the disease versus choice distinction

There is another set of meanings that Szasz attributed to the concepts of bodily and mental illness, namely, that the former is involuntary and the latter is voluntary. In this section, I will argue that it is *these* two meanings, *not* Szasz's claims about lesions or values, that explain his position that mental illnesses are not real illnesses and that clarify what exactly he means by claiming that the concept of mental illness functions to deny agency in others and in general.

Szasz argued that third difference in meaning that distinguishes the concepts of bodily illness from the concept of mental illness is that bodily illnesses are things that happen to us, i.e., they are involuntary, and that mental illnesses are things that we do, i.e., they are voluntary:

"While diseases such as syphilis and tuberculosis are in the nature of *events* and hence can be described without taking cognizance of how men conduct themselves in their social affairs, hysteria and all the other so-called mental illnesses are in the nature of *actions.* They are made to happen [p. 201] by sentient, intelligent human beings and can be understood best, in my opinion, in the framework of games. Mental illnesses thus differ fundamentally from bodily diseases, and resemble, rather, certain moves or tactics in playing games" (1961 p. 200-201 original emphasis).

Szasz explicitly states that the distinction that he is making here is crucial to his argument that mental illness is a myth, i.e. that mental illnesses are not real illnesses[32]. The reason that Szasz takes mental illnesses to be voluntary is because he takes mental illnesses to be learned behaviours[33]. Szasz takes learned behaviours to necessarily entail voluntary behaviour. Indeed, Szasz often summarises his argument that mental illnesses are not real illnesses by contrasting mental illnesses, defined as learned (mis)behaviours, with real illnesses, defined as processes beyond an agent's self-control[34]. Szasz argues that his argument that 'mental illnesses cannot be real illnesses because misbehaviours cannot be real illnesses' is an analytical claim[35]. Two things need to be clarified here on what Szsz means by analytical. First, while Szasz does think that many of the things we call mental illnesses are not real illnesses, his primary position is conceptual, not empirical. Szasz is primarily concerned with arguing that, given how real illnesses are bodily illnesses, and how bodily illnesses and mental illnesses are defined as concepts, mental illnesses cannot *by definition* be real illnesses any more than there can be married bachelors.

---

[32] "The distinction between happening and action is crucial to my argument, not only in this chapter but throughout this book. I have suggested that, in general, we view physicochemical disorders of the body—for example, cancer of the colon—as happenings; and that we view so-called mental illnesses or psychiatric disorders—for example, a hand-washing compulsion—as actions" (1961, p. 154)

[33] "Virtually all behavior with which the psychoanalyst and psychiatrist deal is learned behavior" (1961, p. 153).

[34] "If mental illnesses are diseases of the central nervous system (for example, paresis), then they are diseases of the brain, not the mind; and if they are the names of (mis)conducts … then they are behaviours, not diseases" (Szasz, 1997 p. 76).

[35] "When I assert that (mis)behaviors are not diseases I assert an analytic truth, similar to asserting that bachelors are not married…" (Szasz, 2004 p. 319).

Second, Szasz simply seems to take it for granted that real illnesses are defined as bodily illnesses and that bodily illnesses are defined as involuntary. That bodily illnesses are regarded as involuntary is a fairly common intuition (if not a presumption) to have if we take bodily illnesses to be prototypical diseases, such as cancer, diabetes, and heart disease. It is of course the equation between real illness and all and only bodily illnesses, that most critics of Szasz reject. But at the very least, Szasz seems to be right that, when it comes to uncontroversial cases of diseases, most of these diseases have the feature of being involuntary, where involuntary roughly means that the primary mechanism of the condition (such as metastasis of cancer cells, a failure of the pancreas to produce insulin) are not the kind of things that agents have volitional control over, such as actions, learned behaviours, and aspects of ourselves such as our character. As for his definition of mental illness as voluntary, Szasz seems primarily to base this definition on the fact that historically, when the concept of mental illness came about, the predominant explanation and conception of the things we call mental illnesses *were* of them as learned behaviours according to psychoanalytic theory. But given their general deterministic framework, psychoanalysts regarded learned behaviours as involuntary, to be explained primarily by psychic forces, such as childhood trauma, the id, ego, and superego; as processes largely beyond the control of the agent. Szasz's perspective, however, is that learned behaviours by definition must be voluntary, since learning implies an agent who *does* the learning and is capable of doing otherwise. Szasz therefore sees the concept of mental illness primarily as a *decision* to regard learned behaviours as explainable from an involuntary perspective as opposed to a an agency-based voluntary perspective. The significance of this point is not just that Szasz sees the concept of mental illness implying involuntariness in the particular individual that the concept might be applied to. More importantly, Szasz sees the concept of mental illness, as such, as signifying a prevailing decision in the history of science and medicine to explain goal-directed and learned human behaviour as such in involuntary terms (Szasz, 1996). Consequently, Szasz sees the concept of mental illness as not just denying agency in the people it is applied to, but as denying agency as a possible candidate explanation for human behaviour, and thereby as a denial of agency in everyone. And this, I argue, is the primary reason that Szasz argues that mental illnesses are not illnesses.

This reading of Szasz has explanatory power. It explains why Szasz takes mental illnesses to be distinctly moral in nature, since they are voluntary. It explains why Szasz is equally against non-biological psychiatry since he regards them as equally undermining and denying agency and moral theories as explanations for human behaviour. It also explains what Szasz means by the concept of mental illness functioning to deny agency. Szasz sees the function of the concept of agency as denying agency in the particular people it is applied to, *and* functioning to deny and undermine voluntary explanations as such. His primary opposition to the concept of mental illness is that it represents a rejection of the disease versus choice distinction, where this distinction is being rejected on the basis that all human problems can be explained without reference to agency.

## —4. The disease versus choice distinction as an unanswered question for medicine

The purpose of this fourth and final section is twofold: First, to argue that Szasz's claim that the concept of illness implies involuntariness and mental illness implies voluntariness is contested, and that he therefore takes for granted that medicine ought to maintain a disease versus choice distinction. Second, to argue that whether medicine ought to maintain a disease versus choice distinction remains an unanswered question for medicine, such as in the status of addiction debates. To the extent that such questions remain unanswered, I argue that Szasz's main criticism of modern medicine has not been fully addressed by either his critics or the literature at large.

Thomas Szasz argues that his claim that mental illnesses are not real illnesses follows from the meaning of these two concepts. That is, Szasz argues that he is making an analytical claim when he says that mental illnesses are not real illnesses. Analytical claims, however, are claims which are often uncontroversial with respect to the meaning of the concepts in question. For example, being an unmarried man is simply what the concept of bachelor means. To validate such a statement, we need not observe bachelors in the world, but only look at the definition of the concept 'bachelor.' An argument to support the idea of analytic concepts is that it is impossible to observe bachelors in order to validate the definition of the term. How would we even know what to look for in the world? To get our search for bachelors started, we already need a definition of what a bachelor is. But once we have the definition, we already know what a bachelor is because we have a criterion by which to identify them in the world. Thus, so the argument goes, we have no need to look at existing bachelors, as that could not add anything to the definition that we already have. Although a classic example, the concept of bachelor is an uncontroversial concept for which there is little grounds to dispute its definition. However, while the notions of 'involuntariness' and 'voluntariness' are sometimes associated with the concept of illness and mental illness respectively, they do not at all refer to the definition or central meaning of these two concepts in the way that being an unmarried man is central to the concept of 'bachelor'. Indeed, whether illnesses necessarily should imply voluntariness appears to precisely what is controversial and in dispute.

But while Szasz's claim that mental illnesses are not illnesses being an analytical claim is too strong, it is a legitimate question to ask whether diseases should be regarded as involuntary and whether medicine should primarily concern itself with involuntary conditions and not moral problems of living. How then are diseases primarily regarded as meaning and how does that meaning relate to involuntariness? In other words, is there a disease versus morality distinction?

There appears to be quite a wide consensus that medical concepts such as disease and illness ought to be defined by some kind of negative consequence, such as harm, death, suffering, impairment, or disability (Scadding, 1967). Much debate has been had about how to distinguish the negative consequences that we should regard as medical diseases from normal suffering. The most popular solutions to distinguishing medical from normal suffering is to add some kind of

dysfunction criterion to the concept of disease which points to a deviation in normal functioning (Boorse, 1975; Wakefield, 1992). Such a criterion is supposed to distinguish medical diseases from supposed normal suffering. Such a definition also makes clear that diseases need not be strictly biological in their aetiology. Such a definition was also used to argue for the validity of mental illnesses as real illness and that psychiatry is a legitimate branch of medicine (Spitzer, 1978).

So where does this conception of disease leave the disease versus choice distinction? If we take the view that diseases are involuntary as an aetiological conception of disease, then this broader conception of disease as defined by their consequences is necessarily a rejection of defining diseases primarily as being involuntary in nature. However, such a rejection does not necessarily mean that disorders are involuntary and non-moral, but rather that disorders can be voluntary and moral in nature. Indeed, various proponents of this broad conception of disease as medical disorder argue this when it comes to various conditions. Lious Charland, John Sadler, Peter Zachar, and Nancy Potter all argue in their own ways argue that a complete disease versus morality distinction is untenable (Charland, 2004, 2006, 2010; Sadler, 2005; Zachar & Potter, 2010a and 2010b; Zachary, 2011; Potter, 2013). Sadler and Zachar in particular rely on the broader conception of disease to argue that a strict disease versus morality distinction is not possible because voluntary and moral processes can involve harmful dysfunctions.

However, sometimes the concept of disease is used to refer to involuntary conditions and also to imply that medicine ought to be primarily concerned with conditions of an involuntary nature. Take for example the history of the debate regarding the medical status of addiction (Leshner, 1997; Ferentzy, 2001; Chavigny, 2013.) Much of addiction's medical status rested historically on the argument that addiction is a disease precisely because it is not a moral problems, and it was a disease and not a moral problem because addictive behaviour was argued to be involuntary. Being 'involuntary' gave addiction its non-moral and medical status. Even when it is accepted that addiction would still be a medical disorder because of its consequences, choice models of addiction (in part) face barriers to acceptance over concerns that they will undermine addiction's disease status. Some diseases therefore appear to be 'more equal than others'. Moreover, even those who argue that a complete disease versus morality distinction is impossible to maintain, such as Sadler and Zachar, do not mean that they regard moral problems of living to be the legitimate concern of medicine (Sadler, 2005; Zachar, 2013). Rather, Sadler and Zachar still maintain that conditions with moral components must have certain 'indispensable non-moral features', such as a dysfunction criterion. That is, conditions can only be both moral and medical if they also meet the general conditions for a being a medical disorder on 'non-moral grounds'. But there is no reason to think that dysfunctions cannot be voluntary or moral in nature, such as learned maladaptive behaviours. To assume otherwise is to assume that a condition cannot be both a disease and a moral problem in the same respect or to assume that, to the extent it can be both in the same respect, medicine *ought* to be concerned with primarily non-moral diseases.

In either case, there appears to be a dilemma of sorts that the disease versus choice and morality distinction poses for medicine's conflicting images of itself. On the one hand, medicine has this image of itself as primarily concerned with diseases and not moral problems of living. On the other hand, medicine has tried to adopt this view that diseases and medical conditions need not strictly be biological in nature but can also be psychological or social in nature, and that therefore diseases should be regarded as being equal in their medical status as far as their aetiology is concerned. But if some conditions involve underlying voluntary processes and moral dimensions that are harmful, yet diseases are supposed to be equal with respect to their aetiology, then these two self-images of medicine are necessarily in tension with one another. This tension plays itself out in debates over the status of conditions such as addiction and cluster B personality disorders. It appears that medicine should have to part ways either with the view that all disorders are equal with respect to aetiology, or with the disease versus choice and morality distinction. In this sense, Szasz's criticism that medicine ought to maintain a disease versus choice distinction is unresolved.

**References**


American Psychiatric Association (1952). *Diagnostic and Statistical Manual. Mental Disorders. First Edition.* Washington, DC: American Psychiatric Association.

American Psychiatric Association (1968). *Diagnostic and Statistical Manual. Mental Disorders. Second Edition.* Washington, DC: American Psychiatric Association.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual. Mental Disorders. Third Edition.* Washington, DC: American Psychiatric Association.

Boorse, C. 1975. On the distinction between disease and illness. *Philosophy and Public Affairs* 5: p. 49–68.

Charland, L.C. 2004. Character: Moral treatment and the personality disorders. In The philosophy of psychiatry: A companion, ed. J. Radden. Oxford: Oxford University Press. p. 64–78.

Charland, L.C. 2006. The moral character of the DSM IV cluster B personality disorders. Journal of Personality Disorders 20:2, p. 116–125.

Charland, L. C. 2007. Benevolent theory: moral treatment at the York Retreat. *History of Psychiatry*, 18:1, p. 61-80.

Charland, L. C. 2008. A moral line in the sand: Alexander Crichton and Philippe Pinel on the psychopathology of the passions. In L. C. Charland & P. Zachar (Eds.), *Fact and value in emotion*. Amsterdam, The Netherlands: John Benjamin Press.

Charland, L.C. 2010. Medical or moral kinds? Moving beyond a false dichotomy. Philosophy, Psychiatry, & Psychology 17:2, p. 119–125.

Charland, L. C. 2011. Moral undertow and the passions: two challenges for contemporary emotion regulation. *Emotion Review*, 3:1, p. 83-91.

Chavigny, K. A. 2013. "An Army of Reformed Drunkards and Clergymen": The Medicalization of Habitual Drunkenness, 1857–1910. *Journal of the History of Medicine and Allied Sciences* 69:3, p. 383-425.

Dammann, E. J. 1997. "The myth of mental illness:" continuing controversies and their implications for mental health professionals" Clinical Psychology Review, 17:7, p. 733-756.

Engelhardt, T. Jr. 2004. Mental illness as a myth: a methodological re-interpretation. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler, p. 365-371.

Ferentzy, P. 2001. From sin to disease: differences and similarities between past and current conceptions of chronic drunkenness. Contemporary Drug Problems 28, p. 363-390.

Fulford, K. W. M. 1989. Moral Theory and Medical Practice. CUP, Cambridge.

Fulford, K. W. M. 2004. Value-based medicine: Thomas Szasz's legacy to twenty-first century psychiatry. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Goosens, W. K. 1980. Values, Health, and Medicine. *Philosophy of Science* 47:1, p. 100-115.

Heather, N. 2013. Is alcohol addiction usefully called a disease? *Philosophy, Psychology, & Psychiatry* 20:4, p. 321-324.

Kendell, R. E. 1975. The concept of disease and its implications for psychiatry. *British Journal of Psychiatry* 127, p. 305-315.

Kendell, R. E. 2004. The Myth of Mental Illness. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler p. 29-48.

Leshner, A. I. 1997. Addiction is a brain disease, and it matters. *Science*, *278*, p. 45-47.

Loughlin, M. & Miles, A. 2015. Psychiatry, objectivity, and realism about value. In Alternative perspectives on psychiatric validation: DSM, ICD, RDoC and Beyond, ed. P. Zachar, D. ST. Stoyanov, M. Aragona, and A. Jablensky p. 146-163.

Moss, G. R. 1968. Szasz: review and criticism. *General Review in Psychiatry* 31;2, p. 184-194.

Murphy, D. 2006. *Psychiatry in the scientific image*. Cambridge, MA: MIT Press.

Pearce, S. 2011. Answering the Neo-Szaszian Critique: Are Cluster B Personality Disorders Really So Different? *Philosophy, Psychiatry, & Psychology,* 18:3, p. 203-208.

Potter, N. N. 2013. Moral evaluations and the cluster B personality disorders. *Philosophy, Psychiatry, & Psychology, 20:3, p.* 217–9.

Reimer, M. 2013. Moral disorder in the DSM-IV? The cluster b personality disorders. Philosophy, Psychiatry and Psychology, 20:3, p. 203–15.

Reimer, M., and B. Day. 2013. Affective Dysfunction and the Cluster B Personality Disorders. *Philosophy, Psychiatry, & Psychology,* 20:3, p. 225-229.

Satel, S., & Lilienfeld, S. O. 2014. Addiction and the brain-disease fallacy. *Frontiers in Psychiatry*: Review Article, 4:141, p. 1-11.

Scadding, J. G. 1967. Diagnosis: the clinician and the computer. *Lancet* 2, p. 877-882.

Schuur, R. 2019. Mental Health and Illness: Past Debates and Future Directions. Bloomsbury Companion to Philosophy of Psychiatry, edited by S. Tekin and R. Bluhm. p. 527-541.

Sedgwick, P. 1973. Illness: Mental and Otherwise. *The Hastings Center Studies* 1;3, p. 19-40.

Spitzer, R. L. & Endicott, J. 1978. Medical and mental disorder: Proposed definition and criteria. In: Spitzer, RL.; Klein, DF., editors. Critical Issues in Psychiatric Diagnosis. New York, NY: Raven Press; p. 15-39.

Sadler, J. Z. 2005. *Values and Psychiatric Diagnosis*. Oxford: Oxford University Press.

Sadler, J. Z. 2008. Vice and the Diagnostic Classification of Mental Disorders: A Philosophical Case Conference. *Philosophy, Psychiatry, & Psychology* 15:1, p. 1-17.

Sadler, J. Z. 2013. Values in Psychiatric Diagnosis and Classification. In The Oxford Handbook of Philosophy and Psychiatry Edited by K.W.M. Fulford, Martin Davies, Richard G.T. Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton. Oxford: Oxford University Press. p.

Szasz, T. S. 1960. The Myth of Mental Illness. The American Psychologist 15:2, p. 113-118.

Szasz, T. S. 1961. The uses of naming and the origin of the myth of mental illness. American Psychologist 16:2, p. 59-65.

Szasz, T. S. 1961. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins.

Szasz, T. S. 1974. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins.

Szasz, T. S. 1972. Bad habits are not diseases: A refutation of the claim that alcoholism is a disease, *The Lancet 2, p.* 83—84.

Szasz. T. S. 1987. Insanity: the idea and its consequences. New York: John Wiley.

Szasz, T. 1991. Diagnoses are not diseases. The Lancet 338, p. 1574-1576.

Szasz, T. 1996. A Brief History of Medicine's War on Responsibility. Journal of Clinical Epidemiology, 49:6, p. 609-613.

Szasz, T. S. 1997. Mental illness is *still* a myth. *Review of Existential Psychology & Psychiatry* 23, p. 70-80.

Szasz, T. S. 2000. Mind, Brain, and the problem of responsibility. Society 37:4, p. 34-37.

Szasz, T. S. 2001. Mental illness: psychiatry's phlogiston. *Journal of Medical Ethics* 27: p. 297–301.

Szasz, T. S. 2004a. Reply to Kendell. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2004b. Reply to Pies. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2004c. Reply to Bentall. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2004d. Reply to Slovenko. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2004e. Reply to Engelhardt. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2004f. Reply to Percival. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Szasz, T. S. 2006. Defining disease: The gold standard of disease versus the fiat standard of diagnosis. *The Independent Review: A Journal of Political Economy* 10, p. 325-336.

Szasz, T. S. 2010. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins. 50 anniversary edition.

Szasz, T. S. 2011. The myth of mental illness: 50 years later. Psychiatrist 35, p. 181-184.

Pies, R. 1979. On myths and countermyths. *Archives General Psychiatry* 36:2, p. 139-144.

Pies, R. 2004. Moving Beyond the "Myth" of Mental Illness. Szasz Under Fire: the psychiatric abolitionist faces his critics, edited by J. A. Schaler.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundaries between biological facts and social values. *American Psychologist* 47:3, p. 373–388.

Zachar, P. 2011. The clinical nature of personality disorders: Answering the neo-Szazian critique. Philosophy, Psychiatry, & Psychology, 18:3, p. 191–202.

Zachar, P., and N. N. Potter. 2010a. Personality disorders: moral or medical kinds—or both? Philosophy Psychiatry & Psychology 17:2, p. 107–117.
Zachar, P., and N. N. Potter. 2010b. Valid Moral Appraisals and Valid Personality Disorders. Philosophy Psychiatry & Psychology 17:2, p. 131–142.

**The Puzzle of Addiction, Knowledge of the Good, and Cognitive Integrity**

The occurrence of some harm is a necessary condition for what defines a mental disorder (Spitzer & Endicott, 1978). No one would 'choose' to endure such harm, all else being equal. This is because we have as a general 'rule of thumb' that individuals do not engage in behaviour[36] that is harmful to them if: (1) their behaviour is voluntary, (2) they know the behaviour is harmful, and (3) a positive alternative to that harmful behaviour is available (Pickard, 2015). Indeed, many medical disorders involve behaviours or mental processes that are involuntary, such as bodily shaking in Parkinson's. But there are also mental disorders that involve processes that are voluntary.

For example, while the ritualistic behaviour in Obsessive Compulsive Disorder (OCD) is described as a 'compulsion' and therefore as 'involuntary', it is also widely recognised that such behaviour is not involuntary in the same sense that the bodily shaking in Parkinson's is involuntary. Rather, the ritualistic behaviour in OCD, and compulsive behaviour in general, is understood to be at least partly a voluntary behaviour. Indeed, ritualistic behaviour in OCD is understood to be partly an unconscious attempt to alleviate some underlying anxiety that the person is experiencing[37]. The identification of a 'harmful yet voluntary behaviour' is in itself not a medical mystery because, often, either one or more of the three conditions in our 'rule of thumb' are not met. The individual engaged in ritualistic behaviours, such as excessive hand washing, often knows that their behaviour is harmful. Still, they may not *fully know* that they engage in such behaviour in order to alleviate some deeper anxiety. But even with such a fuller insight, the continued engagement in ritualistic behaviour makes complete sense if no viable alternative to alleviating their intense anxiety is known and has been developed to replace their ritualistic behaviour[38]. While there remains much to be explained of how such ritualistic behaviour occurs, we do have plausible explanations for why people can engage in harmful and voluntary behaviour, such as lacking the (full) knowledge of harm or (full) knowledge of viable alternatives. There is no 'puzzle' of how such behaviour is possible. But there are some mental disorders where the harm of the behaviour in question is so extreme and where the viable alternative appears to be so obvious that it tests our 'rule of thumb' assumptions. The best example of this is addiction.

Addiction is in part defined by persistence in using drugs despite the severe negative consequences for doing drugs, such as financial ruin, psychological distress, familial breakdown, physical deterioration, and even death (APA, 2013). The drug usage of addicts also has many of the hallmarks of voluntary behaviour, such as planning. If continued drug use results in harmful consequences, then how to prevent such harm appears to be obvious for the addict, namely, to stop using drugs. Yet addicts continue to use drugs despite their severe negative consequences and the existence of an available alternative behaviour. This is the 'puzzle of addiction', the puzzle of explaining why addicts continue using drugs despite the severe negative consequences.

---

[36] I also refer to non-behaviour processes, but I will refer primarily to involuntary and voluntary behaviours in this paper.

[37] For example, exposure-based therapies, one of the treatments for OCD, are based on the assumption that individuals with OCD can in principle choose to not engage in a ritualistic behaviour (Olatunji, *et* al. 2013).

[38] Ritualistic behaviour shows that our simple involuntary versus voluntary distinction fails to capture such phenomena.

Solving the puzzle of addiction is important not merely because we want to develop treatments for addiction. The puzzle of addiction also raises deeper questions. First, how can addictive behaviour be both voluntary and harmful? Second, if such behaviour is voluntary, then why does knowledge of its harm appear to be insufficient to stopping such behaviour? Answers to these kinds of questions can serve as the basis for different models for how we can explain, treat, and thereby empower patients as agents within the wider context of medicine as such.

The 'Orthodox' solution to the puzzle of addiction is that the appearance of voluntary behaviour in addiction is merely that, an appearance (see Miller and Chappell, 1991). This solution claims that, in reality, addictive behaviour is involuntary because intense cravings and withdrawal symptoms mean that the addict cannot resist further drug use. Addicts therefore do not stop using drugs because they don't want to, but because they can't and so they don't stop using drugs. On the Orthodox view, we empower addicts by getting them to accept that their addictive behaviour is involuntary, which saves them from inevitably failing to control an uncontrollable behaviour. Such acceptance would then help them realise that they need external and professional help to prevent such behaviour, such as by taking medications that curb cravings.

But the Orthodox view of addiction has been challenged by evidence and arguments that addictive behaviour is voluntary (see Pickard, 2018). As a consequence of such criticisms, the Orthodox view has been criticised for disempowering addicts by undermining and underutilising the role of choice and incentives in addiction. But if addictive behaviour is voluntary, the puzzle of addiction remains: why do addicts keep using drugs despite severe negative consequences?

Gene Heyman has argued that, in order to explain how harmful behaviour such as addictive behaviour can be voluntary and how to change such behaviour, we need to rethink the assumptions that prevent us from making sense of such behaviour and offer a positive alternative to those assumptions (Heyman, 2009). Heyman argues that what gave rise to the assumption that harmful behaviour, such as addictive behaviour, is involuntary is the underlying view that voluntary behaviour is necessarily rational with respect to acting in our best interest. I will call this view 'homo economicus'. Heyman argues that we need to reject homo economicus and replace it with an alternative view of voluntary behaviour. He argues that voluntary behaviour is 'naturally biased' towards acting irrationally against our best interest, a view that I will call 'homo irrationalis'. On the view of homo irrationalis, addicts keep using drugs because their behaviour is biased to viewing the costs and benefits of continued drug use from a short term temporal framework, which Heyman calls 'the local choice framework'. The consequences of such actions only appear as harmful from a longer term perspective, 'the global choice framework'. For Heyman, addicts keep using drugs because they are biased to frame their actions from the local choice framework.

From Heyman's view of homo irrationalis, we can empower anyone engaged in voluntary and harmful behaviour, such as addicts, by getting them to do two things. First, to get them to accept that all human beings have a natural bias to frame their actions from the local choice framework. Second, to 'counteract' such a bias by promoting behaviours that are in the person's long term best interest and are incompatible with the short term benefits of drug use. In the case of addicts, this could include 'private rules' for conduct as can be found in religions that help 'save ourselves from ourselves' and our natural bias towards acting against our best interest.

The purpose of this paper is to criticise and offer an alternative theoretical framework to Heyman's analysis of and solution to the puzzle of addiction from the shared perspective that addictive behaviour is both voluntary and contrary to the best interests of the addict.

I will criticise Heyman's natural bias solution and argue that the underlying assumption that should really be questioned is the view that the global choice framework ought to be sufficient for knowing what is and acting in one's best interest. I argue that the global choice framework is not sufficient for knowing what is in our best interest nor for acting towards it even when we do know what is in our best interest. What is required to act in our best interest is not merely knowledge of what is in our best interest, but epistemic virtues that enable us to act on that knowledge. I will in particular draw on the work by Gena Gorlin on the epistemic virtue of 'Cognitive integrity' to make this argument. On this view, we can fail to act in our best interest even when we have knowledge of what is in our best interest because acting on such knowledge requires Cognitive integrity. My solution to the puzzle of addiction therefore draws on Heyman's insight that having a long term perspective is necessary for both gaining knowledge of what is in our best interest and acting on such knowledge, but without viewing such a long term perspective as sufficient for acting on such knowledge. This solution therefore avoids the problems facing Heyman's natural bias solution.

This paper is divided into three parts. In Part one, I present Heyman's criticism of the Orthodox view and his own solution to the puzzle of addiction. In Part one, section one, I present Heyman's argument that it is the assumption of homo economicus that justifies the Orthodox view. I also present here Heyman's alternative view of voluntary behaviour, homo irrationalis. In Part one, section two, I present Heyman's arguments against the Orthodox view, which involves his definition of voluntary behaviour as behaviour that changes in response to incentives and data that addictive behaviour changes in response to incentives. In Part one, section three, I present Heyman's own solution to the puzzle of addiction, which involves his distinction between the local and global choice frameworks and his appeal to a natural bias to the local choice framework.

In Part two, I criticise Heyman's analysis of the Orthodox view and his solution to the puzzle of addiction. In Part two, section one, I distinguish between a strong and weak version of the Orthodox view. I argue that only the strong version of the Orthodox view accepts Heyman's definition of voluntary behaviour. I criticise the weak version for not clearly defining involuntary behaviour. In Part two, section two, I criticise Heyman's appeal to a natural bias in his solution to the puzzle of addiction by arguing that his idea of a natural bias is vague and in many ways resembles the weak Orthodox view and faces many of the same problems. In Part two, section three, I argue that the underlying assumption that should be questioned is the view that the global choice framework ought to be sufficient for knowing what is and acting in one's best interest.

In Part three I present my solution to the puzzle of addiction. In Part three, section one, I draw on the work of Hanna Pickard on denial to argue against Heyman's assumption that the global choice framework ought to be sufficient for knowing what is and acting in our best interest. In Part three, section two, I draw on the work of Gena Gorlin on Cognitive integrity to argue for the necessity of wanting to know in order to gain knowledge of and act in our best interest. In Part three, section three, I present the failure to develop Cognitive integrity as an alternative solution to Heyman's natural bias solution for why we can act contrary to our best interest despite knowing it.

## —1.0 Heyman's analysis of and solution to the puzzle of addiction

Gene Heyman's solution to the puzzle of addiction is not only significant because it offers a novel solution to explaining why addicts keep using drugs despite the severe negative consequences. His solution is also significant because it presents a whole new framework for making sense of how harmful behaviour can be voluntary as such. His solution to the puzzle of addiction therefore constitutes not just a positive theory about addiction, but furthermore a positive theory about the nature of voluntary behaviour. Heyman's positive theory of addiction and voluntary behaviour, however, are a direct response to his analysis of past theories of addiction and voluntary behaviour. In particular, Heyman argues that the view that addictive behaviour is involuntary is due to the view that voluntary behaviour as such is necessarily rational with respect to acting in our best interest. It is this view, Heyman argues, that gives rise to the view that addictive behaviour, because it is harmful, is necessarily involuntary. Heyman's project is not merely to show that the view that addictive behaviour is involuntary is itself false, but that the underlying view of voluntary behaviour that gives rise to it is false. Presenting Heyman's analysis of this view of voluntary behaviour is essential for two reasons. First, because Heyman's own theory of addiction and voluntary behaviour is based on his analysis and criticism of what this other view of voluntary behaviour consists of. Second, because I disagree with Heyman's analysis of what gives rise to the view that addictive behaviour is involuntary. This disagreement is crucial since it also forms that basis for my later criticism to Heyman's solution, and it forms the basis for my alternative framework for making sense of voluntary harmful behaviour.  Before presenting Heyman's analysis of what gives rise to the view that addictive behaviour is involuntary, I want to clarify what we should call such a position. The view that addictive behaviour is involuntary is historically associated with what has been called the 'disease model' and later the 'brain disease model' of addiction. However, there are disagreements about whether the disease model necessarily depends on the view that addictive behaviour is involuntary (see Segal, 2013). Moreover, the argument that addictive behaviour is not involuntary can be wrongly taken as an argument that addiction is therefore not a medical disorder at all. Because of the ambiguous relation between the concepts of 'disease', 'disease model' and 'involuntariness', some scholars have instead called the view that addictive behaviour is involuntary the 'Orthodox view' (Pickard, 2016). The 'Orthodox view' is merely meant to refer to the fact that, for many decades, the predominant explanation for why addicts keep using drugs despite its negative consequences is that addictive behaviour is somehow involuntary. Calling this position the 'Orthodox view' is meant to delimit the discussion on whether or not the view that addictive behaviour is involuntary is true and leaves aside questions of whether or not the disease model of addiction is necessarily committed to this view or whether diseases are necessarily involuntary. As we will see, Heyman views both the disease model of addiction and calling addiction a disease as necessarily entailing a commitment to the view that addictive behaviour is involuntary. Heyman's project is therefore also to argue that the disease model of addiction is wrong and that addiction is not a disease, though he does call addiction a disorder that warrants societal and clinical intervention. For the purposes of this paper I am only concerned with Heyman's project to reject and replace the Orthodox view of addiction.

## —1.1 Heyman's analysis of the Orthodox view of addiction

Gene Heyman starts his analysis of the Orthodox view, the view that addictive behaviour is involuntary, with an observation of how it is often justified by appealing to scientific evidence, such as claims that addiction is associated with various genetic predispositions or that addiction 'changes the brain' (Leshner, 1997). However, Heyman argues that such appeals to empirical scientific evidence are not primaries, meaning that they do not, and indeed cannot, demonstrate to us whether addictive behaviour is involuntary or not. Why not? To demonstrate this point, Heyman points out that there are many behaviours that are associated with certain genes and brain changes that we do no regard as involuntary, such as religious beliefs (Heyman, 2009 p. 94). Why, then, are the brain changes and genetic associations linked with addictive behaviour regarded as empirical scientific evidence that addictive behaviour is involuntary whereas brain changes and genetic associations linked with religious belief are not? Heyman argues that the reason why one set of data is regarded as evidence for involuntary behaviour whereas the other set of data is not regarded as evidence for involuntary behaviour is not because the data sets are fundamentally different in kind. Rather, Heyman argues that such data sets do not, and indeed cannot, on their own show that one behaviour is involuntary and the other behaviour is involuntary. This is because our very concepts of what it means for a behaviour to be voluntary or involuntary do not come from modern genetics or neuroscience. Indeed, there are no areas in the brain that 'light up' as 'involuntary' on an fMRI scan, and genes associated with certain behaviour are not labelled as 'involuntary' in genetic sequencing experiments. Rather, Heyman argues, our concepts of voluntary and involuntary behaviour are based on behavioural criteria and observations, that is, distinctions and assumptions that we make about behaviour that we can actually observe. In fact, modern genetics and neuroscience are based on establishing certain associational and causal relations between phenotype behaviours and the data that these modern sciences provide us. You could therefore not even begin to look for empirical scientific evidence that a particular behaviour is involuntary or not if you didn't already have some basic conception of what you were looking for, i.e., behavioural criteria for voluntary and involuntary behaviour. This is certainly not to say that the scientific empirical evidence presented in favour of the view that addictive behaviour is involuntary is therefore necessarily question-begging. Rather, it is simply to say that, when it comes to determining whether any behaviour is voluntary or involuntary, we cannot start by simply looking at such kind of evidence. Instead, we have to start by looking at the behavioural criteria we use to distinguish between involuntary and voluntary behaviour[39].

This raises the question of how we do distinguish involuntary from voluntary behaviours? Heyman first provides his analysis of what he calls the 'traditional rule' for making this distinction, before developing his own criteria for making this distinction (Heyman, 2009 p. 98).

---

[39] "…the debate about the nature of addiction has been framed as a biological issue, yet the biological data have not helped solve it. The reason is that the criteria for deciding whether an activity is voluntary are behavioral. We do not look at people's genes to determine if they are engaged in a voluntary or involuntary act, we look at their behaviour. Similarly, we do not look at their brains to decide if their actions are voluntary or not. This is not to deny that there are biological underpinnings to the distinction between voluntary and involuntary acts. Rather, the point is that the distinction rests on criteria that precede what we have learned about the brain" (Heyman, 2009 p. 97-98).

Heyman argues that addiction came to be regarded as an involuntary condition because addictive behaviour was regarded as self-destructive. He traces this view back to 17th century English clergymen (Warner, 1994). As Heyman explains: "The reasoning behind the clergy's belief that they were dealing with a disease is that their "sick" parishioners kept drinking despite drinking related problems" (Heyman, 2009 p. 98). According to Heyman, the so-called 'traditional rule' behind the clergy's reasoning is the assumption that voluntary behaviour cannot be overall self-destructive and that therefore self-destructive behaviour must be involuntary behaviour.

It is this 'traditional rule' which Heyman argues led to the view that addictive behaviour is involuntary behaviour because addictive behaviour is contrary to our best interests[40]. Heyman points to quotes of contemporary proponent of the Orthodox view to argue that they too view addictive behaviour as involuntary because it is self-destructive[41][42][43].I will call this view *homo economicus*, the view that voluntary behaviour is necessarily rational with respect to acting in our best interest (where rationality is defined as calculating the best cost benefit analysis between two or more choices to achieve the best overall outcome). But Heyman points out that this is by no means the only way to view the nature of voluntary behaviour. Indeed, Heyman argues that there is another view that is in opposition to the view of homo economicus. As Heyman explains:

"Western culture offers two contradictory visions of voluntary behavior. The seventeenth-century preachers and twenty-first-century addiction scientists take the widely held position that choices are fundamentally rational, hence no one willingly engages in self-destructive ends. This has been formalized in economics, where the assumption is that choices do not simply avoid self-harm but maximize benefits and minimize costs. In contrast to this view, the arts often portray individuals as knowingly, willingly, and persistently pursuing self-destructive ends" (Heyman, 2009, p. 101).

---

[40] ''The seventeenth-century account of alcoholism is not based on biology but on widely shared understandings of voluntary behavior. The fundamental assumption is that individuals make choices that are in their best interests. Since addiction is self-destructive, the logical implication is that addicts cannot be voluntarily choosing to use drugs. Since the symptoms of diseases are involuntary, then addiction must be a disease. In other words, medical evidence did not turn alcoholism into a disease, but rather the assumption that voluntary behavior is not self-destructive turned alcoholism into a disease" (Heyman, 2009 p. 99).

[41] "A metaphorical switch in the brain seems to be thrown as a result of prolonged drug use. Initially, drug use is a voluntary behavior, but when that switch is thrown, the individual moves into the state of addiction characterized by compulsive drug seeking and use" (Leshner 1997, p. 46).

[42] "Rarely overtly stated but clearly central to the concept of a disease is a victim state. As a victim, the afflicted has no control over the progression of the disease if left untreated. In the disease concept of alcoholism (and drug addiction), the cardinal feature is *loss of control* over the use of alcohol… The loss of control, which can actually be inherited, is the sine qua non for alcoholism (and drug addiction) as qualifying for the disease state. The loss of control signifies a victim that reflects an alteration of brain function" (Miller and Chappel, 1991 p. 197 emphasis by Heyman, 2009 p. 100).

[43] "Today, as in the seventeenth century, it is assumed that individuals do not repeatedly engage in self-destructive behavior unless they are compelled to do so. Addiction is a pattern of persistent self-destructive behavior, hence it is involuntary, and as diseases are involuntary, then addiction is a disease. This is perfectly logical, given the assumption that voluntary behavior is necessarily not self-destructive. Again, the issue of whether addiction is a disease depends on the understanding of voluntary behavior" (Heyman 2009 p. 100).

The view here is that human voluntary behaviour is not only capable of being irrational with respect to acting in our best interest, but that voluntary behaviour is more often than not irrational with respect to acting in our best interest. Heyman draws on both ancient literature, mythologies, and religions which depict human behaviour as often flawed with regards to acting in our own best interest. Heyman argues that in such literature we see human beings as both knowingly and willingly acting contrary to their own best interest. Like homo economicus, Heyman does not provide a formal definition of what this view amounts to. But like homo economicus, the way that Heyman presents this view tells us something about what he thinks this view amounts to.

First of all, this view is not a negative view, that is, it is not merely the rejection of homo economicus. It is not merely the position that the view that voluntary behaviour is necessarily rational with respect to acting in our best interest is false. Rather, it is a positive view of the nature of voluntary behaviour, a positive view which is also a rejection of homo economicus. However, what precisely this positive view of voluntary behaviour amounts to is never clearly spelled out by Heyman. Since Heyman sometimes presents this view as being in contrast with or contradictory to homo economicus, one way to read it would be that it is defined as the exact opposite view, namely, that voluntary behaviour is biased to irrationality with respect to acting in our best interest.

However, while Heyman does define homo economicus in terms of a necessary relation between voluntary behaviour and rationality, he does not define this opposing view in terms of necessity. Rather, Heyman describes this view's conception of voluntary behaviour as being often irrational or inclined towards irrationality with respect to acting in our best interest. As we will see later, Heyman also describes it as including a 'natural bias' towards certain kinds of irrationality.

I will come back to Heyman's characterisation of this view and argue that it has serious problems as the basis for a framework to solve the puzzle of addiction because it is not clearly defined. I will call this view of voluntary behaviour 'homo irrationalis', the view that voluntary behaviour is biased towards or inclined to irrationality with respect to acting our best interest.

Heyman introduces this view for two reasons. First, the very existence of this view provides some initial basis to consider whether or not the homo economicus view of voluntary behaviour is true. Second, if homo economicus is not true, homo irrationalis provides an alternative framework for explaining harmful but voluntary behaviour such as in addiction. As Heyman himself explains:

"One of the lessons of literature then (and, I would argue, everyday experience as well), is that voluntary acts are often self-destructive. This implies that self-destructive drug use is not the proper criterion for determining that someone is compulsively using drugs. Likewise, it implies that rationality is not the correct criterion for distinguishing between voluntary and involuntary acts. We need some other approach" (Heyman, 2009 p. 102).

I will now turn to Heyman's criticism of the Orthodox view before presenting his own solution.

—1.2 Heyman's criticism of the Orthodox view of addiction

Heyman argues that whether the Orthodox view of addiction is true comes down to whether its underlying view of voluntary behaviour, homo economicus, is true. But Heyman argues that we first need a definition of voluntary behaviour in order to test whether the Orthodox view is true.

Heyman argues, however, that there is a problem when it comes to providing a definition of voluntary behaviour since voluntary behaviour is often defined as rational behaviour. However, whether voluntary behaviour is necessarily rational is precisely what is at issue. Heyman therefore argues that we need a definition of voluntary behaviour that is 'neutral' on the issue of whether or not voluntary behaviour is or is not necessarily rational with respect to acting in our best interest.

Heyman starts off his process of distinguishing between involuntary and voluntary behaviour by pointing to a well-established and empirically verified distinction between two kinds of behaviours, namely, behaviours that are elicited in response to their settings, such as external stimuli, and behaviours that are elicited and mediated in response to their consequences, such as rewards and costs (Heyman, 2009 p. 104). In the former category we have behaviours such as reflexes and tics that are relatively immune from rewards or costs. In the latter category we have any kind of behaviour that can be brought about, changed, or brought to a halt in response to incentives, such as rewards and costs. Heyman uses this distinction between these two kinds of well-established behaviours to roughly distinguish between involuntary and voluntary behaviours. While some behaviours might be elicited by both stimuli and incentives, the general distinction between these two kinds of behaviours are sufficient for Heyman's purposes. As Heyman explains:

"These contrasts provide a rule for determining whether an activity is voluntary or involuntary. Voluntary activities vary as a function of costs, benefits, the opinions of others, cultural values, and the myriad of other factors that influence decisions. Involuntary activities vary little or not at all as a function of the factors the influence decisions. Thus, we can test whether drug use in addicts is voluntary by testing whether it is brought to a halt by the factors that influence decisions" (Heyman, 2009 p. 103-104).

Heyman does offer one qualification for voluntary behaviours, namely, that the behaviours must be feasible in light of alternative options. For example, food is necessary for survival, and so most people cannot reasonably be expected to forgo eating food since there is no feasible alternative to it. Therefore, eating can be regarded as functionally involuntary for the purposes of changing behaviour that we want to stop people from engaging in. Heyman's definition of voluntary behaviour is therefore: any behaviour of a person that (1) is subject to change in response to incentives, and (2) for which feasible alternatives to that behaviour are possible. Moreover, this definition is 'silent' on the issue of rationality, which for Heyman is essential since he regards having rationality as a criterion of voluntary behaviour to lead to the problematic assumption that self-destructive behaviour must be involuntary, which he rejects. As he explains:

"…we need a definition of voluntariness that is silent about rationality. The idea that behaviors can be evaluated in terms of the degree to which they vary as a function of their consequences does just this. According to this rule, voluntary activities vary systematically as a function of their consequences, where the consequences include benefits, costs, and values. In contrast, involuntary activities are elicited by preceding stimuli (e.g., urges) and are influenced little or not at all by their consequences" (Heyman, 2009 p. 104).

Heyman's definition of voluntary behaviour therefore does not only allow us to test whether addictive behaviour is involuntary or not. Heyman's definition of voluntary behaviour provides us with a definition to assess whether or not the homo economicus view is true. If voluntary behaviour always results in the best possible outcomes over time then the homo economicus view is true. If voluntary behaviour often results in outcomes that are overall bad with respect to our best interest, then the homo economicus view is false and an alternative view of voluntary behaviour is required to explain how voluntary behaviour can so often be harmful. As we will see, Heyman will argue that homo economicus is proven to be false by the empirical data on voluntary behaviour, and that homo irrationalis provides a framework to explain harmful voluntary behaviour.

Heyman draws on three kinds of empirical evidence to argue that addictive behaviour is voluntary: autobiographical evidence, remission rates, and contingency management programmes (CMP's). Autobiographical accounts of addiction show that many addicts stop using drugs when faced with financial ruin, family break down, or a challenge to their identity (such as having to become a drug dealer in order to continue supporting their habit).

Moreover, Heyman argues that such cases are the norm, not the exception, for addicts. Data on remission rates for addiction show that most addicts recover without professional help and do so by their late 20's to early 30's. But Heyman takes the autobiographical evidence to be unscientific since it cannot be replicated. Furthermore, he argues that the remission data is inferential, not causal, so he does not rest his case on them either. Rather, Heyman draws on CMP's to demonstrate that addictive behaviour is indeed voluntary. CMP's are programmes where addicts are given modest vouchers (for food, other material necessities, or recreational activities) in return for 'clean' urine samples. And the longer that addicts adhere to the programme, the more that the monetary worth of their vouchers increases. Not only do these programmes have high rates of abstinence while addicts are in treatment, which can reach up to 70% for five weeks, but rates of abstinence also continue to improve even after the programme ends (Heyman, 2009).

Addictive behaviour therefore does appear to change in response to incentives. Heyman argues that, because addictive behaviour changes in response to incentives, that addictive behaviour is therefore voluntary behaviour. Moreover, he argues that both the Orthodox view of addiction and its underlying homo economicus view of voluntary behaviour are false. I will turn now to Heyman's explanation for how addictive behaviour can be both voluntary and self-destructive.

## —1.3 Heyman's solution: voluntary behaviour as biased towards self-destruction

So far, Heyman has argued that the Orthodox view of addiction fails to solve the puzzle of addiction and that the underlying view of homo economicus is false because voluntary behaviour is not necessarily rational with respect to acting in our best interest.

But this leaves unresolved the puzzle of addiction: we need a positive explanation for how behaviour can be both voluntary and harmful, with respect to acting in our best interest. That is, we need an alternative view of voluntary behaviour to solve the puzzle of addiction[44].

Heyman's solution to the puzzle of addiction and his alternative view of voluntary behaviour that it rests on has two components. First, a distinction between a local and global choice framework to frame the cost-benefit analysis of our choices. Second, a 'natural bias' as an explanation for why we tend to frame the cost and benefits of our actions from a local choice framework. But before presenting these two components, Heyman first introduces a thought experiment to illustrate three principles of choice that form the basis of his overall solution.

In this thought experiment, you have to choose between eating out at a Chinese or an Italian restaurant each night. The initial conditions are that you have a preference for Chinese food over Italian food, but your preference for Chinese food both decreases faster the more you eat it and increases faster the longer you don't eat it. Thus, if you choose Chinese food over Italian food more than a couple of times in a row, you will come to prefer Italian food, but if you switch to eating Italian food over time your preference for Chinese food will return more quickly.

The question is, if we set overall gastronomic enjoyment as our goal, how should we behave to maximise gastronomic enjoyment? When Heyman asked his students this question they came up with two kinds of strategies. The first and by far the most common strategy was to eat at which ever restaurant one had the highest preference for each night. The second strategy was to choose restaurants to maximise gastronomic enjoyment over time. These strategies resulted in markedly different sets of choice patterns and overall gastronomic enjoyment.

Heyman uses the term 'equilibrium' to refer to the total pattern of individual choices that emerges from one of the two choice strategies over time. The first strategy, choosing which ever meal is valued most in the moment, produces an equilibrium pattern where many more Chinese meals were consumed than Italian meals over time. Now while the second strategy, maximising overall gastronomic enjoyment, also produced an equilibrium pattern where Chinese meals were consumed more often than Italian meals in total, this strategy resulted in more consumption of Italian meals and less consumption of the Chinese meals relative to the first strategy.

---

[44] "Although there are many sources of support for the vision of voluntary behaviour that has been presented in this chapter, it is incomplete. There is no explanation of how behaviour can be both voluntary and self-destructive. To be sure, the empirical and logical relations introduced in this and previous chapters support the statement that "addicts voluntarily choose to use drugs in a self-destructive manner." But nothing has been said as to how this comes about. Theories of motivation and choice typically predict good if not optimal outcomes. Thus, the conclusion that addiction is self-destructive yet voluntary calls for an explanation" (Heyman, 2009 p 114).

Moreover, the second strategy theoretically produces higher levels of gastronomic enjoyment over time. This is due to the habitation and dis-habituation function of the Chinese meals, whereby its value decreased faster the more it was consumed, and increased faster the less it was consumed. The first strategy therefore resulted in less overall gastronomic enjoyment.

Heyman argues that these two choice strategies show three principles of choice.

First, choices and preferences are dynamic: we choose things based on what we prefer and our choosing them changes our preferences over time. For example, our choosing the Chinese meal in itself changes the value of future Chinese meals over time, namely, choosing them decreases their future value and not choosing them increases their future value.

Second, preferences shift as a function of temporal framing: pleasure now versus long-term enjoyment changes how we choose. Heyman calls this these the local versus global choice frameworks, respectively. The local and global choice frameworks refer to the temporal horizon by which we can frame our choices, with the local choice framework referring to smaller units of time in comparison to the global choice framework, where the unit of time is larger.

What this means is that whether or not a behaviour is rational is a function of what the ultimate goal is that is pursued and how both the behaviour and goal are framed temporally. A behaviour that therefore appears to be irrational with respect to acting in our long term best interest could turn out to be rational with respect to acting in our best interest in the short term.

Third, we always choose the best option given our temporal framework. Heyman does not really explain this 'rule' and it is presented more as a postulate than based on any observations.

Heyman argues that these three principles, taken by themselves, are uncontroversial. Yet, when taken together, they result in two surprising outcomes. First, the local choice framework resulted in a significant reduction in overall food enjoyment over time. This shows that taking a local choice framework can explain why behaviour can be voluntary and sub-optimal. Second, Heyman points out that the local choice framework dominates in guiding behaviour in both humans and other species (Heyman, 2009 p. 123). As Heyman explains,

"As the local equilibrium approximates what is typically found in both laboratory and natural settings, the message is that some degree of overconsumption usually accompanies voluntary actions" (Heyman 2009 p. 124).

This conclusion gives evidence against the homo economicus view that voluntary behaviour is necessarily rational with respect to acting in our best interest. As Heyman explains:

"…choice can stabilise at a suboptimal level of benefits, that suboptimal yet voluntary outcomes involve overconsumption of at least one of the options, and that contingencies that guide choice are ambiguous. These conclusions are at odds with the assumption that voluntary actions are guided by rationality" (Heyman, 2009 p. 124).

Moreover, Heyman points out that this kind of behaviour would be anomalous under the homo economicus view of voluntary behaviour and would be explained away as a single instance of bias or irrationality, rather than a 'principle' or 'fundamental feature' of voluntary behaviour[45]. As Heyman explains: "As long as choices are made from the local perspective, and this is usually the perspective that people take, the favoured good will be consumed excessively" (2009, p. 135).

Heyman argues that behaviours can be both voluntary and harmful because the perceived costs and benefits of any behaviour are a function of how we frame it. In particular, the local versus global choice framework distinction shows that the rationality of a behaviour is a function from how it is framed. While a behaviour can be regarded as irrational from the perspective of the global choice framework, the very same behaviour may turn out to be completely rational from a local choice framework. Heyman argues that this often appears to be the case in addicts. For example, if one is a homeless addict, then 'sleeping rough' on the streets may be more tolerable if one uses drugs than if one doesn't. From a local choice perspective, the choice to continue using drugs can often appear to be a rational choice, even though the very same behaviour is irrational from the perspective of one's long term best interests. Heyman argues that addicts, in a sense, continue using drugs because they are stuck in a local choice framework regarding how they frame their choices. On this view, addicts will only stop using drugs if and when the costs of using drugs starts to outweigh their benefits from the perspective of the local choice framework. Both the epidemiological data on remission rates and evidence from contingency management programmes predict this, Heyman claims. We should treat addiction, then, by thinking about how we can increase the costs of drug use and increase drug-free alternative behaviour from the local choice framework. This includes incentivising behaviour that is in the addict's best interest from the global choice framework which in itself makes such behaviour more rewarding in the future.

Heyman does acknowledge, however, that the calculations involved in the global choice framework may not be equal to the local choice framework, and may in themselves exact a cost. However, Heyman does not factor such costs into his analysis, treating them as irrelevant.

As Heyman explains:

"[The global choice framework] may exact a psychologist cost and the plan itself may exact cognitive costs, but in terms of physical effort or monetary expense (recall, it is assumed that all meals have the same price), the two approaches are identical. Thus, how the options were framed made a difference, all else being the same" (Heyman, 2009 p. 122)[46].

---

[45] "Both economics and behavioural biology imply that a theory of excess is not needed. According to economics, individuals and firms tend to end up at the global equilibrium. In cases where this is not achieved, the explanation is that some sort of mistake or bias is at play, not principle. According to biologists, organisms maximise fitness. For both disciplines, excessive maladaptive consumption patterns are a kind of irrationality or accident; phenomena that do not fit the standard analyses. In contrasts, in the local/global analysis, *excess is a fundamental feature of voluntary behaviour*" (Heyman, 2009 p. 134 emphasis added).

[46] "Some calculations suggest that under most circumstances the local equilibrium is not that different from the global equilibrium (e.g., Heyman, 1982, 1983), and the additional benefits provided by the global equilibrium may be somewhat smaller than indicated by the graphs in this book. This is because global bookkeeping is necessarily more complicated, and the costs associated with the complexity of the calculations were not included in the graphs (as it is not clear how to do so)" (Heyman, 2009 p. 129)

But if we can treat the 'operational' costs of the two frameworks as being more or less equal, why then would anyone choose the local choice framework over the global choice framework if the latter is better? Indeed, Heyman claims that we should expect the global choice framework to 'win out' over time, but yet it doesn't[47]. Heyman claims that cases where the global choice framework *is* achieved point to why the local choice framework is dominant.

Experiments show that only humans are capable of discovering the global choice framework 'on their own' (Heyman, 2009 p. 138). All other animals need to be indirectly taught to approximate behaviour consistent with the global choice framework by manipulating how choice options are presented. Heyman argues that this species difference might have to do with our capacity to abstractly reflect on our choices over time. That is, from one perspective the global choice framework refers to our capacity to abstractly form long term goals and engage in complex reasoning on how to achieve them. This is distinct from choices within the local choice framework, which we can perceive directly, such as 'this meal', as opposed to our 'overall food enjoyment' as such. Heyman then claims that this difference might explain why the local choice framework dominates and why we have a 'natural bias' towards it (Heyman, 2009 p. 172).

One way of reading such a 'natural bias' is that, precisely because operating a global choice framework to assess one's action requires cognitive effort, it is simply easier to engage in the local choice framework. That is, the local choice framework could be regarded as a kind of 'default option' since it requires no effort to engage in whereas the global choice framework is far from the default option but rather requires real work to achieve. Moreover, the difference in the ability of human beings to achieve the global choice framework on their own, when compared to non-human animals, may point to the role of abstract thinking, which requires more effort. Thus, a 'natural bias' might refer to the local choice framework just being easier to adopt[48]. The local choice framework may also just be easier to grasp than the global choice framework[49]. But this runs contrary to Heyman claiming that we can treat the operational costs of the two frameworks as equal. Moreover, Heyman speaks of this 'natural bias' not merely as the need to minimise effort, but as a positive inclination that human beings might have towards the local choice framework. Indeed, Heyman is also drawing on the homo irrationalis view, the view that voluntary behaviour is not only capable of irrationality but has some kind of natural bias towards such irrationality. This suggests that 'natural bias' is a distinct explanation from merely cognitive effort.

---

[47] "In almost all ordinary situations, the local equilibrium provides a lower rate of overall benefits than does the global equilibrium. Often the difference are quite small. But as we are almost always in a setting that offers more than one alternative, the differences should add up. Thus, everyday events as well as highly seductive opportunities lead local choice astray. These observations say that the local perspective should give way to the global perspective. Learning is like natural selection. Thus, if local and global choice compete for the same niche, global choice should win. Nevertheless, in most studies the local equilibrium prevails" (Heyman, 2009 p. 135).

[48] "The species differences are intriguing. They suggest that the capacity to reflect upon the options is one of the factors that distinguishes global choice from local choice…. [G]lobal choice is more cognitively demanding than local choice" (Heyman, 2009 p. 138).

[49] "The options in the local choice are concrete and correspond to how things look. Local choice involves items that have clear physical outlines and activities that are easy to name…. In contrast, the aggregates of global choice have no naturally occurring boundaries, but are abstractions…Put another way, local choice corresponds to the natural fracture lines of perception; global choice does not….Thus, local choice persists despite its drawbacks because it is simpler and the options of local choice are consistent with how thing like and their customer labels" (Heyman, 2009 p. 138-139)

If we assume there is a natural propensity to the local choice framework, this can explain why destructive behaviours can occur, such as in addiction. But Heyman argues that this explanation now raises the opposite puzzle: why aren't there more addicts? As Heyman explains:

"Research supports this darker image of voluntary behaviour. In labs and in natural settings, choice proportions approximated the local, not the global equilibrium. This in turn suggests that addiction and other forms of excess should be quite common. However, societies cannot function well if their members are so easily seduced by "specious" rewards. These considerations suggest that there is a role for measures that protect people from themselves…" (Heyman, 2009 p. 140).

According to Heyman, various factors protect against addiction and help approximate the global choice framework. Individual differences in cognition, such as the ability to abstract, appear to protect against addiction, as does marriage, which could promote long term planning. But there are many counter-examples to both of these protective factors: cognitively inclined addicts, the sober but 'less'-cognitively inclined, addicts who are married, and sober singles.

Heyman points to a paper by Drazen Prelec and Richard Herrnstein, titled *Preferences or Principles* (Prelec and Herrnstein, 1991), which argues that most people do not follow a rational decision making process whereby they weigh the costs and benefits of their actions long term. Indeed, calculating the short and long term costs and benefits of every action isn't just something that most people do not do, it is not even practical, given our limited time and cognitive capacity. Rather, most people appear to apply 'private rules' for how to conduct themselves, such as wearing a seat belt because 'that's what a safe driver does'. These private or prudential rules appear to also be mediated by social norms and identity, such as views about what a good person does and how we view ourselves. Such prudential rules and social identities appear to 'substitute' rationality, according to Heyman, and help approximate the global choice framework.

Heyman also points to the negative correlation between religiosity and addiction, and the effectiveness of AA as a treatment for addiction, to argue that these prudential rules and social identities help prevent and treat addiction by providing an approximation to the global choice framework by substituting the global choice perspective on rationality that we lack (Heyman, 2009 p. 166-173). In other words, prudential norms provide rules that help protect us from ourselves, i.e., they protect us from our natural bias to frame our choice from the local choice framework.

Heyman argues that, to explain destructive but voluntary behaviour, we should view voluntary behaviour as 'naturally biased' towards irrationality with respect to acting in our best interest. Such a view does provide a plausible solution to the puzzle of addiction. I will now turn to my criticisms of Heyman's solution in Part two before presenting my own solution in Part three.

## —2.0 Criticism of Heyman's analysis of and solution to the puzzle of addiction

In Part two I criticise Heyman's analysis of and solution to the puzzle of addiction. I argue that Heyman's analysis of the Orthodox view of addiction is only partially true. I argue that there are in fact two versions of the Orthodox view of addiction, a strong version that accepts Heyman's definition of involuntary behaviour and a weak version that rejects it. The strong version is rightly rejected by Heyman's arguments, but the weak version is not. But the main problem with the weak Orthodox view is that it is explanatorily weak and circular, since it isn't clear what it means by 'involuntary' behaviour. I suggest that what also motivates both versions of the Orthodox view is that they provide a social justification to treat addicts under the purview of medicine.

I also criticise Heyman's claim that what causes the puzzle of addiction is the view that voluntary behaviour is necessarily rational with respect to acting in our best interest, a view which I have called homo economicus. I argue that none of the players in this debate that Heyman ascribes this view to actually holds this view. The 17th century English clergymen had a view of voluntary behaviour closer to Heyman's own view of homo irrationalis, the view that voluntary behaviour is naturally biased towards acting irrationally with respect to our best interest.

Indeed, my main criticism of Heyman's solution is that it is not fundamentally different from the weak version of the Orthodox view. Heyman's solution consists of first distinguishing between the local and global choice frameworks to show that how we frame our actions determines whether they appear rational or not with respect to our best interest. But given how Heyman conceives of this distinction, it creates the problem of why we don't always operate from the global instead of the local choice framework if we have access to the former. His solution to this problem is to argue that humans have a natural bias towards the local choice framework in order to explain why we don't always operate from the global choice framework. But this natural bias solution is functionally the same as the weak Orthodox View because it argues that we are not equally free to choose to frame our actions from the global choice framework. But it is not clear what it means for us to have this natural bias to framing our actions, it is hard to see what it would take to disprove this view, and it takes for granted that this bias is natural instead of learned and thereby more voluntary, which is what a solution to the puzzle of addiction is supposed to show.

But Heyman is correct that there is underlying assumption in the puzzle of addiction debate that needs to be questioned. This is the assumption that the global choice framework ought to be sufficient for knowing what is and acting in one's best interest. This assumption is explicit in homo economicus since it holds that the global choice framework ought to produce optimal behaviours. This assumption is implicit is Heyman's homo irrationalis since Heyman regards the global choice framework as such as being theoretically sufficient to act in one's best interest. The fact that we do not act in our best interest is evidence for Heyman that some kind of natural bias must explain why we don't always act from the global choice framework. In Part three I will criticise the assumption that the global choice framework ought to be sufficient to both know what is and act in our best interest by drawing on the work of Hanna Pickard on the role of denial in addiction. I will then draw on the work of Gena Gorlin on Cognitive integrity as a framework to explain how we can act contrary to our best interest even when we have knowledge of it.

—2.1 Criticism of Heyman's analysis of the Orthodox view

In this section I will criticise Heyman's analysis of the Orthodox view. I will argue that Heyman's analysis that the Orthodox view is based on a view of voluntary behaviour that I call homo economicus is wrong. Moreover, I argue that the Orthodox view consists of a strong and a weak version and that both have problems with solving the puzzle of addiction. I also argue that part of the motivation for the Orthodox views is that they address the 'social' puzzle of addiction.

Heyman's criticism of the Orthodox view rests on his definition of voluntary behaviour as behaviour that changes in response to incentives (Heyman, 2009 p. 102-103). Proponents of the Orthodox view have either agreed with or rejecting Heyman's definition of voluntary behaviour.

The first kind of response is to agree with Heyman's definition of involuntary behaviour but to argue that addictive behaviour is nonetheless involuntary (see Miller and Chappel, 1991). There are generally three ways that proponents argue that addictive behaviour is still involuntary.

First, in response to the epidemiological evidence that most addicts do not remain addicts, some argue that addicts do not change their behaviour in response to incentives but that they instead go into 'spontaneous recovery' (Miller and Chappel, 1991). Indeed, the phenomenon of spontaneous recovery is real, such as cancerous tumours inexplicably shrinking.

Second, it has been argued that not all drug users are engaged in involuntary behaviour but rather that only a subset of them exhibit involuntary behaviour (Jellinek, 1960a; Jellinek, 1960b). Those persons labeled as 'addicts' whose behaviour does change in response to incentives might be regarded as having a disorder but not as being engaged in involuntary behaviour, whereas those addicts whose behaviour does not change in response to incentives can be described as being engaged in involuntary behaviour. But this response has fallen out of favour since it conflicts with claims that addiction is heavily under-diagnosed (Leshner, 1997).

Third, instead of distinguishing between kinds of addicts, as in the last case, it has also been argued that there are 'stages' or 'phases' of addiction *within* the addict, where addictive behaviour can be voluntary and respond to incentive in the chronic stage and become involuntary in the acute stage of the condition (Leshner, 1997). Accordingly, when we see addicts being able to change their behaviour in response to incentives, this does not necessarily mean that addictive behaviour is always voluntary, it just shows that it is voluntary during the chronic stages.

These three ways of trying to retain the position that addictive behaviour is involuntary do so because they want to retain the explanatory value of involuntariness to solve the puzzle of addiction. Indeed, these kinds of distinctions could be completely legitimate since they are made in other areas of medicine as well. But the main problem with these kinds of responses is that these various distinctions in order to retain the explanatory value of involuntariness can actually come at the expense of the explanatory power of the Orthodox view. Any one of these responses could be used to counter any criticism of the Orthodox view to the point where the position becomes unfalsifiable. For example, if addicts who appear to be in an acute stage of drug use still manage to resist using drugs and recover, the Orthodox view could just argue that the acute stage ended and that their addictive behaviour went from involuntary to voluntary, but that the 'involuntary stage' could still return. Such responses undermine the view's explanatory power.

I call this first response the 'strong version' of the Orthodox view because it has a clear and narrow conception of what it means for addictive behaviour to be involuntary. Its advantage is that it provides a clear explanation for why addicts keep using drugs, namely, addictive behaviour does not change in response to incentives. Its two major disadvantages are that the strong version of the Orthodox view is either clearly false, because addictive behaviours do change in response to incentives (Heyman, 2009), or it becomes explanatorily weaker when its response to claims that it is false are met with an over-reliance on the kind of distinctions mentioned above.

The second kind of response to Heyman is that his definition of involuntary behaviour fails to capture what the Orthodox view actually means by addictive behaviour being involuntary[50].

These proponents of the Orthodox view of addiction will argue that Heyman's definition of involuntary behaviour is too narrow, namely, that behaviour simply not responding to incentives is too literal of a meaning for when addictive behaviour is described as being 'involuntary'. Gabriel Segal, for example, criticises Heyman by arguing that no addiction expert seriously compares the shaking in Parkinson's with the behaviour of addicts as being involuntary in the same sense[51]. Alan Leshner's *has* compared the behaviour of addicts with the shaking in Parkinson's disease[52], but Leshner also acknowledges that studies demonstrate that addicts *can* change their behaviour in response to incentives[53]. Leshner argues that addicts do have significant responsibility for taking control of their addictive behaviour, and that an addict is therefore not a 'hapless victim':

"…the recognition that addiction is a brain disease does not mean that the addict is simply a hapless victim. Addiction begins with the voluntary behavior of using drugs, and addicts must participate in and take some significant responsibility for their recovery. Thus, having this brain disease does not absolve the addict of responsibility for his or her behavior, but it does explain why an addict cannot simply stop using drugs by sheer force of will alone" (Leshner, 2001).

---

[50] "If Heyman just means to point out that drug use, even in addicts, is not like tics, accidents, and automatism, then we can agree with him. Addicts by and large intend to use drugs, and the particular actions that constitute such use—tightening the tourniquet, pressing the syringe—are under their executive control… Do proponents of the disease model claim that the actions of addicts are just like tics or reflexes? If they do, they are mistaken, and Heyman would be right to criticize this kind of claim. But they are more likely invoking a notion of compulsion that…does not require that the behavior fails to be action at all. To say that an action is compelled signifies that there are very significant automatic or involuntary elements involved that compromise the individual's capacity to choose and act differently" (Kennett, 2013 p. 268-269).

[51] "But "compulsive" does not always mean anything so extreme. Even people with obsessive-compulsive disorder can say "no" and resist their "compulsions" to some extent (Abramowitz 2006). In fact, no contemporary theoretical model of addiction offered by neuroscientists in the "disease" camp treats addictive behavior as completely beyond voluntary control, like Parkinson's shaking (Hyman 2007). What, then, do theorists in the camp mean by "compulsive?" Steven Hyman explains (Hyman 2007, p. 2): The term compulsion is imprecise, but at a minimum implies diminished ability to control drug use, even in the face of factors (e.g. illness, failure in life roles, loss of job, arrest) that should motivate cessation of drug use in a rational agent willing and able to exert control over behavior" (Segal 2013, p. 452).

[52] "Over time the addict loses substantial control over his or her initially voluntary behavior, and it becomes compulsive. For many people these behaviors are truly uncontrollable, just like the behavioral expression of any other brain disease. Schizophrenics cannot control their hallucinations and delusions. Parkinson's patients cannot control their trembling" (Leshner, 2001).

[53] "…two large sets of multisite studies have demonstrated the effectiveness of well-delineated outreach strategies in modifying the behaviors of addicted individuals that put them at risk for acquiring the human immunodeficiency virus (HIV), even if they continue to use drugs and do not want to enter treatment. This approach runs counter to the broadly held view that addicts are so incapacitated by drugs that they are unable to modify any of their behaviors" (1997 p. 45).

Leshner argues that the ‘brain disease’ conception of addiction can help us explain why addicts cannot stop using drugs through ‘sheer force of will alone’. While it is true that a ‘sheer force of will alone’ is not sufficient to heal oneself of a brain disease, it does not follow that if a 'sheer force of will alone’ fails to explain why addicts do not stop using drugs that they therefore have a brain disease. Indeed, most problems we face in life cannot be solved through a ‘sheer force of will alone’. Developing one’s character for the better, for example, cannot be done through ‘sheer force of will alone’, but requires practicing self-awareness and moral knowledge.

Either what Leshner means by a brain disease is something narrow like Alzheimers or something broad to mean that any psychological problem has a basis in the brain. Marc Lewis has provided a critique against viewing addiction as a brain disease in a narrow sense that I will not get into here (Lewis, 2018). The broad conception of a brain disease has little explanatory value other than showing us that ‘sheer will’ alone is insufficient to achieve most things in life.

But if Heyman’s definition of involuntary behaviour is too narrow then this also weakens the Orthodox view’s solution to the puzzle of addiction. If involuntary behaviour is defined as behaviours that do not change in response to incentives, then this provides a strong explanation for why addicts keep using drugs despite the costs outweighing the benefits. But if this is not what the Orthodox view means by ‘involuntary’, then its explanatory power diminishes with a broader definition of involuntary behaviour. Indeed, some have argued that if addictive behaviour is involuntary and yet can change in response to incentives then it is not at all clear what it means for addictive behaviour to be involuntary nor therefore how this solve the puzzle of addiction (Heather, 2013). As Hanna Pickard has argued with regards to the Orthodox view,

“The appeal to compulsion understood as irresistible desire is key to the orthodox conception's explanation of persistent use in the face of negative consequences. Suppose that, even if the desire to use is hard to resist, it is not irresistible. Then the question of why use persists in the face of negative consequences remains. For, given the severity of these consequences, the *difficulty* of resisting - as opposed to the *impossibility* of resisting - is not by itself explanatory. We need to know more….[T]he parsimony and power of the orthodox conception to explain the puzzle of addiction depends on an appeal to compulsion understood as irresistible desire. Softening the meaning of compulsion costs the orthodox conception its explanatory force” (Pickard, 2018 p. 10 original emphases).

I call this latter version of the Orthodox view the ‘weak’ version because it has a broader conception of involuntary behaviour than the ‘strong’ view. That is, according to the weak version of the Orthodox view, addictive behaviour can be described as involuntary and yet still change in response to incentives. The advantage of the weak version of the Orthodox view is that it is not clearly false and does not have to rely on various distinctions to address counter-examples, though sometimes the weak version also relies on such distinctions. The disadvantage of the weak version of the Orthodox view is that its conception of involuntary behaviour is so broad it is therefore hard to tell what it means for addictive behaviour to be involuntary, and because of its conceptual vagueness of the term ‘involuntary’ the weak version thereby loses explanatory power.

All of this analysis so far has assumed that the puzzle of addiction and the Orthodox view that seeks to solve it concerns primarily an explanatory problem. But clearly another important factor motivating the Orthodox view is that addicts also pose a kind of social or societal problem.

This social puzzle of addiction has two components. First, how do we as a society get addicts that are not seeking treatment to accept that they have a problem? This can especially be a problem when addicts themselves can be in denial that they even have a problem and/or face social stigma. Saying that addicts have a disease that is involuntary is easy to understand and to accept. Second, how do we get society to accept that addicts have a problem that is worth treating? This question was especially important historically because when addiction was viewed to be partially voluntary it was viewed to be moral in nature, with the implication that addicts did not deserve treatment or at least were not as deserving as the 'sick' (see Warner, 1994). The Orthodox view solves this problem by arguing that addictive behaviour is involuntary and not immoral in nature and therefore worthy of treatment by society under the purview of medicine.

However, solutions to the social puzzle of addiction have also had their own problems. While regarding addictive behaviour as involuntary (i.e., a symptom of a disease) may help addicts accept that they have a problem and seek help, it may also hinder their recovery. Indeed, some data seems to suggest that addicts who believe in the disease model of addiction are more likely to relapse than those who do not (Miller, et al. 1996). This makes sense. Addicts are less likely to resist the urge to use drugs if they believe that the desire to use them is irresistible. Proponents of the Orthodox view resolve this dilemma by claiming that addicts do not have voluntary control of their behaviour to consume drugs but do have voluntary control to seek help to stop using drugs:

"Even more paradoxical is the fact that, although the alcoholic is not at fault for having the disease of alcoholism, personal responsibility is the cornerstone in the process of recovery. There is a volitional component to recovery—to seek and accept treatment for the involitional component, i.e., the loss of control over alcohol or drug use. In order to recover from the diseases of alcoholism and drug addiction, outside help is often necessary to strengthen the volition to maintain abstinence and suppress the loss of control" (Miller and Chappel, 1991 p. 203-204).

Here again we have ambiguity of whether addictive behaviour is truly involuntary since it turns out that addicts can 'suppress the loss of control' over their addictive behaviour and urges.

A similar problem faces the Orthodox view's solution to social rejection of addicts by fighting stigma (Racine, et al. 2015). While more public funding and support has gone into addiction research and treatment, critics of the Orthodox view argue that much of this funding has gone into research and treatment that presumes that addictive behaviour is involuntary, such as medications to curb cravings, despite research showing that addictive behaviour is voluntary and that treatment programmes that presume this, like contingency management programmes, have much better results than craving-curbing medications (see Satel and Lilienfeld, 2014). Satel and Lilienfeld have called this a 'Faustian bargain', since calling addiction a brain disease has increased funding for addiction but has misdirected funding away from the treatments that do work, because they reject the Orthodox view, and towards treatments that are not as effective.

The Orthodox view has also been shown to reduce the moral stigma of addiction but it seems to have increased other forms of stigma (Racine, *et al*., 2015). Critics of the Orthodox view have argued that it relies on a simplistic view of how to address stigma (Frank and Nagel, 2017).

Indeed, it is worth noting that throughout the history of addiction there have been times when the involuntariness of addiction has been highlighted for non-explanatory ends which are then corrected for when those non-explanatory ends were met (see Warner, 1994). One example was the inebriety reforms in the UK in the 19th century, which argued that 'habitual drunkenness' must be regulated because it posed a threat to public health (Johnstone, 2001). But once the inebriety reforms were implemented, proponents of the movement wished to emphasise that calling addiction a disease could be misunderstood as meaning that addicts were incapable of controlling their behaviours, which ran contrary to treatments which in part relied on the agency of addicts to achieve recovery[54]. Another example of this kind of course correction is Alan Leshner's 'rebranding' of addiction as a disease to a brain disease. Leshner is quite explicit in his reflections on his role as the head of the National Institute of Drug Abuse that part of calling addiction a brain disease was motivated by attempts to get more support for addiction research and treatment, especially in terms of more funding from the United States Congress (Leshner, 2010). But once such funding was secured, and the brain disease label was criticised as undermining the role of choice in addiction recovery, Leshner clarified that addiction being a brain disease does not mean that addiction does not involve choice, both in the condition and its recovery (Maasing, 2004)[55].

This kind of dynamic points to two phenomena with regards to the puzzle of addiction. First, the puzzle of addiction is not merely an explanatory puzzle but also a social puzzle of justifying and supporting the treatment of problems like addiction that some might think are not worth treating if it turns out that addiction is not a medical condition. Second, that there is the potential that a solution to the social puzzle of addiction can come at the expense of a solution to the the explanatory puzzle of addiction. For example, we may want to argue that addiction is a disease primarily because we want to secure access to medical care for addicts even if calling addiction a disease, or a brain disease, interferes with our ability to explain and treat addiction.

But to the extent that the puzzle of addiction is an explanatory puzzle, what causes it? Heyman argues that it is the view that voluntary behaviour is necessarily rational with respect to acting in our best interest and that therefore irrational behaviour with respect to our best interest like addictive behaviour is necessarily involuntary, a view I have called homo economicus.

---

[54] "By disease is popularly understood a state of things for which the diseased person is not responsible, which he cannot alter except by the use of remedies from without, whose action is obscure, and cannot be influenced by exertions of his own. But if, as is unquestionably true, inebriety can be induced by cultivation; if the desire for drink can be increased by indulgence, and self-control diminished by lack of exercise; it is manifest that reverse effects can be produced by *voluntary effort;* and that the desire for drink may be diminished by abstinence, and self-control, like any other faculty, can be strengthened by exercise. It is erroneous and disastrous to inculcate the doctrine that inebriety, once established, is to be accepted with fatalistic resignation, and that the inebriate is not to be encouraged to make any effort to mend his ways. It is more so, since inebriety is in many cases recovered from, in many diminished, and since the cases which recover or amend are those in which the inebriate himself desires and strives for recovery *(1908 Report:* p. 5-6, emphasis added)"" (in Johnstone, 2001 p. 49).

[55] "…Dr. Leshner maintains that his views have been distorted and misinterpreted. Still, he says, he has lately modified his message, giving more recognition to the role of volition in addiction. "Today's version," he says, is that addiction is "a brain disease expressed as compulsive behavior; both its development and the recovery from it depend on the individual's behavior"" (Maasing, 2000).

Heyman's analysis that the view that voluntary behaviour is necessarily rational with respect to acting in our best interest is central to his analysis of and solution to the puzzle of addiction. He argues that this view is false because addictive behaviour shows that behaviour can be both voluntary and contrary to our best interest. Moreover, Heyman argues in favour of an alternative view of voluntary behaviour, namely, that we are naturally biased to frame the costs and benefits of our actions from the local choice framework, i.e., we are irrational with respect to acting in our best interest from the global choice framework, a view I call homo irrationalis.

Whether or not Heyman is correct in his analysis of both the origins of the Orthodox view and the puzzle of addiction is important in assessing his approach to the puzzle of addiction. Heyman attributes the view that I have called homo economicus to both the 17th century English clergymen who first started calling addiction a disease and he attributes it to contemporary addiction experts. But do these two groups actually hold such a view of voluntary behaviour?

Whether contemporary addiction theorists hold the view of homo economicus is hard to ascertain. Heyman quotes a lot of them to show that they view addictive behaviour to be involuntary and that they do so in part because viewing addictive behaviour as involuntary can help us explain why addicts engage in harmful behaviour (Heyman, 2009 p. 100). But this does not necessarily mean that they hold that voluntary behaviour is necessarily rational with respect to acting in our best interest, as Heyman claims. All this means is that they regard involuntary behaviour as the best explanation for why addicts engage in harmful behaviour. As for the 17th century English clergymen, there is no mention in Jessica Warner's article, the historian that Heyman draws on, that they held that voluntary behaviour is necessarily rational with respect to acting in our best interest. On the contrary, these clergymen described all manner of behavioural problems, even swearing, as 'involuntary' (see Warner, 1994). Moreover, these clergymen described all of these 'habits' as diseases and sins at the same time. This is best explained by their belief in the doctrine of original sin, the view that human nature is somehow deficient and prone to self-destructive actions and therefore requires external help to correct. Such a view is much more consistent with Heyman's own view of voluntary behaviour, homo irrationalis.

There is one sense in which Heyman might be right to attribute the view of homo economicus to the 17th century English clergymen, namely, the view of homo irrationalis may itself be due to holding the view of homo economicus as our expectation of voluntary behaviour. That is, if you hold the conditional view that if we truly were free to act on our knowledge and if we truly did know what was good for us, then one might conclude that if human beings consistently act contrary to their best interest that we are therefore not truly free, rational, or knowledgable. Indeed, one kind of argument for original sin is that human beings would not act as destructively as they do if they truly were free to act and knew what was good for them. The concept of original sin, much like an involuntary disease for addiction, provides an explanation for such behaviour.

But not only does Heyman's view of homo irrationalis more closely resemble the underlying assumptions of those of the Orthodox view. Heyman's solution to the puzzle of addiction also resembles the weak version of the Orthodox view. This and other criticisms of Heyman's solution will be presented next. I will then lay out the assumption that gives rise to Heyman's solution, before questioning it and providing my alternative solution in Part three.

—2.2 Criticism of Heyman's solution to the puzzle of addiction

Heyman's solution to the puzzle of addiction consists of two steps. The first step is to introduce the distinction between the local and global choice frameworks as distinct ways to temporally frame our actions in relationship to their consequences. If we assume that an addict is operating from the local choice framework, then their actions do not appear to be irrational after all but may in fact be rational. It is only from the global choice framework that an addict's actions appear to be irrational with respect to their best interest. The addict keeps using drugs despite the severe negative consequences because the framing of their choices prevents them from fully knowing that they are acting contrary to their best interest. The distinction therefore serves as a kind of argument from ignorance rather than an argument from a deficient will as such. On this view then addicts keep using drugs because the costs have not yet started outweighing their benefits from the local choice framework, and addicts will stop using drugs when the costs do stop outweighing their benefits, either by their lives getting worse or by incentivising them to stop by increasing the costs of drug use and increasing the reward of non-drug activities.

But Heyman points out that there is an important difference between humans and non-human animals with regards to the global choice framework. Non-human animals can only approximate the global choice framework in their behaviours, that is, there is no evidence that they can actually temporally extend the framing of their actions. Humans, on the other hand, can and do. In other words, human themselves can choose to frame and reframe their actions temporally. Indeed, this is precisely what we do when we are asked to engage in the restaurant thought experiment. So while incentives can help us to stop using drugs from the local choice framework, we can also 'self-induce' those incentives by adopting the global choice framework, such as by thinking about the severe negative consequences of our actions and how ceasing drug use can lead to long term rewards. But given how Heyman conceives of the global choice framework, this first step to his solution to the puzzle of addiction also introduces new puzzles.

Heyman's conception of the local and global choice frameworks concerns the framing of our actions in relationship to the value of certain options. However, Heyman is agnostic about how certain basics facts and values are known. Indeed, his thought experiment assumes that all the basic facts are known and the main variable is how we frame those same facts, such as the value of cuisines. Heyman therefore equates having access to a certain temporal framework to having access to certain kinds of knowledge that we already seem to have. Moreover, or perhaps because of this, Heyman does not take there to be much difference in the cost and effort in framing our actions from either the local or global choice framework. At most, he argues that the global choice framework might require more work but argues that this cost should be negligible.

But given Heyman's conception of the global choice framework, that he takes knowledge within it for granted and as functionally equal in its operation cost to the local choice framework, Heyman's solution to the puzzle of addiction faces its own puzzle. Given these assumptions, why don't we always frame our actions from the global choice framework? In other words, why would we ever frame the costs and benefits our our actions from the local choice framework when framing the costs and benefits of our actions from the global choice framework is clearly better?

Heyman's solution to this problem is that human beings, while they have access to the global choice framework, have a 'natural bias' towards to the local choice framework. What this means is that we are not equally free to choose the global choice framework over the local choice framework. This does not mean that we are unfree to choose the global choice framework, just less free. In practice this means that our choices are always weighted towards the local choice framework and weighted against the global choice framework. This explains, for Heyman, why we sometimes do choice the local choice framework over the global choice framework. Moreover, it is a crucial part to Heyman's solution to the puzzle of addiction. The first part is that addicts keep using because the costs of using drugs does not outweigh the benefits from the local choice framework. The second part is that the reason why they appear to be stuck in the local choice framework is because all human beings have a natural bias towards such a choice framework. In other words, Heyman's solution to the puzzle of addiction is that we keep using drugs because we have a natural bias to frame our actions from the local choice framework, that is, there is a natural weighting towards framing our actions from a short term temporal horizon.

But Heyman's natural bias solution now creates yet another problem, namely, if we have this natural bias towards the local choice framework, then why don't we see more self-destructive behaviour than one might expect. In other words, why aren't there more addicts? Since the global choice framework is something we are weighted against choosing, his solution needs to come from elsewhere. Heyman argues that most of our actions are not rationally determined, by which he means, we don't weigh the costs and benefits of all of our actions. That would not be practical nor do we appear to be very good at it. Rather, Heyman points out, evidence shows that most of the time people act on private rules rather than rationally assess the benefits of every action, such as the rule that you should always put on your seatbelt when driving. Such rules, Heyman argues, substitute for rationality and approximates what kind of behaviour we would engage in from the global choice framework if we did exercise rationality regarding our actions in our best interest. Heyman also points to institutions and practices that help promote behaviour that approximates behaviour from the global choice framework and protects us from our natural bias towards the local choice framework, such as being married and being employed. Private rules and social practices protect us from ourselves and our natural bias towards the local choice framework.

But while Heyman presents his solution to the puzzle of addiction as distinct from the Orthodox view, it in fact shares many similarities with the weak version of the Orthodox View. While Heyman argues that addictive behaviour is voluntary, he also implies that we are not equally free to choose the global framework over the local choice framework to frame out actions, since we have a natural bias towards the latter. Indeed, Heyman seems to think that we don't really frame our actions from the global choice framework, but rather we substitute other behaviours in order to approximate the how we would act if we did frame our actions global choice framework. 'Natural bias' then functions similarly to 'involuntary behaviour' in the weak Orthodox view as a solution to the puzzle of addiction, by both arguing or implying that we are not fully free to act as we would if we truly were fully rational. The main difference then is that for the weak Orthodox view our will is 'weak' whereas for Heyman how we frame our actions is 'biased'. Their solutions also point to similar treatments, namely, pro-social behaviours to 'save us from ourselves'.

Heyman's solution to the puzzle of addiction is therefore not fundamentally different from the weak Orthodox view in how it tries to solve the puzzle of addiction. It is also not always clear what Heyman means for us or addicts to have a natural bias towards the local choice framework. Moreover, Segal has criticised Heyman for his failure to deal with 'hard cases' where addicts claim to know that the costs of continued drug use outweigh the benefits (Segal, 2013). Heyman can only respond by arguing that such addicts are still framing the costs and benefits of their actions according to the local choice framework because they are naturally biased to do so. Heyman can then claim that if addicts still keep using drugs then this is evidence that the costs have not outweighed their benefits, since if they did, then addicts would stop using drugs. Notice that, much like the Orthodox view, there is always some explanation for why addicts keep using that protects the natural bias solution: we appeal to natural bias to explain why addicts don't stop using and the natural bias is counter-acted when the costs to stop using outweigh the benefits.

—2.3 Heyman's assumption about voluntary behaviour made explicit

Heyman argues that the assumption that voluntary behaviour is necessarily rational with respect to acting in our best interest, homo economicus, underlies the puzzle of addiction and argues that it should be rejected. In its place Heyman proposes the alternative view that voluntary behaviour is naturally biased towards irrationality with respect to acting in our best interest, homo irrationalis. This alternative view is by no means new and may actually be the assumption that underlies earlier arguments for the Orthodox view of addiction. But there does appear to be an underlying assumption that both of the views, homo economicus and homo irrationalis, share. That is the view that something like the global choice framework ought to be sufficient for knowing what is in our best interest, and that knowledge of what is in our best interest ought to be sufficient for us to act in our best interest, i.e., that knowledge of the good is sufficient for good action. Where the views differ is whether we are actually free to act on such knowledge. But they agree that we would act in our best interest if we both had knowledge of it and were free to act on it. You can see this assumption at play in Heyman's work in his restaurant thought experiment. There the highest good, long-term gastronomic fulfilment, is taken as a place-holder for our best interest, and the main question is whether we view it from a short or long term perspective. Indeed, Heyman takes our 'best interest' simply to be a function of what we regard as our highest good from a particular temporal perspective, rather than something we develop from such a perspective. It is for this reason that Heyman takes the global choice framework to be sufficient for knowledge of what is in our best interest since such knowledge is whatever we regard as the highest good. But if our best interest is not simply what we regard as our highest good but refers to knowledge about the self and what is good for us that actually needs to be discovered and developed, then taking a long term perspective on our actions is not sufficient for knowing what is in our best interest. I will argue in Part three that the global choice framework is not sufficient for knowing what is in our best interest. Moreover, I will argue that even when we have knowledge of what is in our best interest that this is not sufficient to produce action that is in our best interest. Rather, I will show that we can fail to act in our best interest even when we have knowledge of it.

**—3.0 Cognitive integrity as an alternative framework for solving the puzzle of addiction**

Heyman's takes the global choice framework to be sufficient for both knowing, and acting in, our best interest. The fact that we often do not act in our best interest leads Heyman to posit a 'natural bias' towards the local choice framework as a solution to the puzzle of addiction. In Part three I will first argue that adopting the global choice framework, even if we had unbiased access to it, would not be sufficient for gaining knowledge of our best interest. I will then argue that having knowledge of our best interest would also not be sufficient for acting in our best interest.

The first argument will be inspired by the work of Hanna Pickard on denial as a solution to the puzzle of addiction. What Pickard's work shows is that taking a long term perspective on ourselves does not simply reveal to us what is in our best interest. It is precisely because such knowledge is an achievement that psychological processes can make us ignorant of such knowledge. Such is the case when it comes to denial. But while denial provides a powerful explanation for why addicts keep using drugs despite the severe negative consequences, it does not explain why addicts return to using drugs once they gain knowledge of those negative consequences. Indeed, acceptance that continued drug use is self-destructive is more often the first to achieving sobriety, not the last. Pickard argues that addicts also require knowledge of *what is in* their best interest, not merely *what is contrary to* it. However, why is knowledge of what is in our best interest insufficient for us to always act in our best interest? How can addicts return to using drugs once they have gained knowledge of their best interest and how to achieve it?

To explain why knowledge of what is in our best interest is not sufficient for acting in our best interest, I will first argue that a *willingness to know* what is in one's interest is a requirement for knowing what is in our best interest. In the same way that knowledge of what is in our best interest is not a given, neither is one's motivation to gain it. I draw here on the concept of Cognitive integrity developed by my colleague Gena Gorlin (Gorlin and Schuur, 2019). Cognitive integrity refers to both a mental state of actively wanting to gain knowledge and to the epistemic character trait of being disposed to seeking such a mental state. Knowledge of one's best interest is complex since it refers to the integration of all of one's interests and its continual development. Therefore, having and developing knowledge of one's best interest requires cognitive integrity.

But once such knowledge is gained, Cognitive integrity is also required to maintain such knowledge *and* to act on it. Failing to act on such knowledge can happen when we default on our willingness to know, or by subverting our minds to non-cognitive goals, such as through denial. This solution to the puzzle of addiction therefore does not depend either on ignorance or a weak will. We can act contrary to our best interest by not maintaining and developing the epistemic virtues required to know and act in our best interest. This solution differs from Heyman's solution in that it recognises the necessity of a long term perspective to act in one's best interest without relying on the presence of a 'natural bias' and the disadvantage it poses. This solution builds on Pickard's denial-solution by showing how our motivation to gain and act on knowledge is required to act in our best interest, even when we already have knowledge of our best interest. This solution also draws on 'Cognitive integrity' as framework to solve the puzzle of addiction.

## —3.1 Hanna Pickard's work on denial as a solution to the puzzle of addiction

Heyman has pointed out that human beings are the only species capable of discovering the global choice framework without external assistance (Heyman, 2009 p. 138). But Hanna Pickard's work on addiction in human beings and rodents suggests a much deeper difference between humans and non-human animals. Rodents are incapable of having knowledge of the causal relation between drug use and their negative consequences, and therefore there is no puzzle for why they keep using drugs (Pickard & Amhed, 2016). The puzzle of addiction only occurs when knowledge that continuing drug use is contrary to one's best interest is actually available. Heyman assumes that the global choice framework ought to be sufficient to having such knowledge, and the fact that we can act contrary to our best interest despite having access to such a framing suggests to him a 'natural bias' to the local choice framework. But Pickard points out that human beings, who can view their best interest from a long term perspective, can face challenges in gaining knowledge of what is in their best interest and whether certain actions support it.

Pickard distinguishes between two ways that we gain knowledge of the causal relation between our actions and their negative consequences (Pickard & Ahmed, 2016). The first concerns causal relations that are generalisable because they have been established through science. The second concerns causal relations that are generalisable because they have been established through individual inferences. Examples of the first kind are cases such as establishing that smoking tobacco products causes cancer. While some short and long term consequences of smoking tobacco products have been known for as long as they have been consumed, such as stained teeth, the fact that they cause a heightened risk of cancer took many years to scientifically establish. Examples of the second kind include gathering inferences that apply to one's own situation as an individual. A rather simple example is that one might be finding out that consuming certain kinds of drugs leads to specific side effects, such as headaches. A more complicated example might be establishing a causal relation between one's alcohol consumption and the deterioration of one's marriage (Pickard, 2016). If alcohol consumption is causally contributing to the deterioration of your marriage, then this might be harder to establish than it may first appear. If, for example, you cease consuming alcohol for a period of time then one's marriage does not automatically improve. Moreover, the poor state of one's marriage may motivate one to either return to or keep consuming alcohol at high levels.

Pickard's main point in establishing these two kinds of ways of gaining causal knowledge is to show that gaining such knowledge is by no means a straightforward process (Pickard, 2016). While such knowledge may seem obvious once a causal relationship is established, or seems obvious from the perspective of external observers, gaining such knowledge in the first place is a complex process that requires both the motivation, patience, cognitive tools, and trial and error to establish. In other words, gaining knowledge of the causal relations between our actions and their negative consequences can be a major cognitive achievement. To the extent that such an achievement gives us knowledge of what is in our best interest, such knowledge does not reveal itself by simply taking a long term view on our best interest, contrary to Heyman. Rather, taking a long term perspective is necessary but not sufficient for gaining knowledge of our best interest.

However, it is precisely because gaining causal knowledge can be an achievement that gaining such knowledge can be derailed by various cognitive processes, such as denial. Hanna Pickard argues that denial can explain why some addicts keep using drugs despite the severe negative consequences (Pickard, 2016). Pickard mentions that denial is a widely recognised phenomenon associated with addiction. However, she points out that denial is rarely addressed within addiction research (Pickard, 2016 p. 278). This might be because thinking of denial as a solution to the puzzle of addiction does not fit within the Orthodox view of addiction being a disease. To the extent that denial is recognised within the Orthodox view, it is the denial that addictive behaviour is involuntary or that addicts need help. But Pickard suggests that denial can explain in certain cases why addicts keep using drugs as such. She argues that addiction can have many causes and therefore that denial does not explain the puzzle of addiction in all cases. But for our purposes, her use of denial as a solution to the puzzle of addiction for some addicts is worth exploring because it points to the role of choice and knowledge in addiction in general.

Pickard defines denial as a kind of motivated irrational belief and as a psychological defence mechanism (Pickard, 2016 p. 285). The primary object of denial that concerns Pickard is the denial of the causal relation between the addict's actions and their negative consequences. But in principle, denial can also be directed at other objects of knowledge. There may be many reasons for an addict to 'be in denial' that their drug use results in severe negative consequences. For example, acknowledging that one's actions have resulted in severe negative consequences can bring about intense emotions, such as shame and guilt of the harm one has caused others and the harm one has caused one's self. Accepting responsibility for such harms can also be very distressing and denial provides a defence against experiencing such emotional distress. One may also be in denial because one does not want to acknowledge some of the reasons why one might be consuming drugs at high levels and prolonged periods in the first place. For example, one may be using drugs to self-medicate because one has experienced traumas that one does not want to face since even acknowledging, let alone processing, such trauma can be very painful. Drug usage may be a way that such traumas can be medicated without facing them. Accepting that one has a drug problem may entail acknowledging that one is using for reasons that are painful to accept. One may also be in denial because accepting that you have a drug problem may entail accepting that one needs to stop using drugs, despite drugs having some limited benefits in meeting certain needs (such as momentary relief from psychological distress). Denial in addiction may therefore be motivated, because accepting that you have a drug problem means that you will likely have to endure more suffering in the short term by addressing the reasons for using drugs and rebuilding the life you had. Simply put, there are many reasons to be in denial.

But while denial seems to be able to explain why addicts keep using drugs despite the severe negative consequences, it does not explain why addicts return to using once they have accepted that continued drug use is contrary to their best interest. Pickard argues that merely accepting that continued drug use is bad for you in not sufficient for recovery. One also requires positive knowledge of what is in your best interest as an alternative to using drugs (Pickard and Ahmed, 2016 p. 37). But even when such knowledge is gained, addicts are still capable of acting contrary to their best interest. Why is such knowledge insufficient for acting in our best interest?

—3.2 Cognitive integrity as an epistemic virtue and its role in knowledge acquisition and action

Pickard's work shows that knowledge of what is in our best interest is not guaranteed by our capacity to take a long term perspective on our actions, i.e., it is necessary but not sufficient. Pickard's use of denial as an explanation for why addicts keep using drugs despite the severe negative consequence also points to the importance of epistemic motivation in lacking or gaining knowledge of what is in our best interest. But how do we explain why addicts can return to using drugs once they have accepted that continued drug use is overall ('globally') bad for them? There are still cases where people appear to act contrary to their best interest despite having some knowledge, however imperfect, of what is in their better interest and how to achieve it.

I suggest that, if we conceive of denial as an epistemic vice, then an alternative epistemic virtue is required to achieve knowledge of what is in our best interest *and* to act on it. But the opposite of denial in not acceptance, since passive acceptance alone is not sufficient to achieve knowledge. The opposite of denial is a willingness to actively engage with reality and develop knowledge of it. This is the epistemic virtue of 'Cognitive integrity' as developed by my colleague Gena Gorlin (Gorlin and Schuur, 2019). I will show how such an epistemic virtue is necessary, not merely for developing knowledge of what is in our best interest, but also, to act on knowledge of what is in our best interest. A failure to foster such an epistemic virtue can explain how we can act contrary to our best interest despite having knowledge of which actions are in our best interest.

Most scholars that study the nature of choice and action have focused on our capacity to choose amongst a number of physical actions. But other scholars in the past have also conceived of the locus of choice as the choice to deliberate. Alexander of Aphrodisias, for example, argued that the choice to deliberate precedes our choice of over physical actions because even choosing to engage in a physical action requires deliberation (Sharples, 2007). The conception of the choice to deliberate is the choice to engage in 'reality-oriented cognition'. Cognition comes from the Latin to cognise, 'to get to know, to recognise'. By reality here I mean both external reality (which is accessed through extrospection) and reality of the self (which is accessed through introspection). To say 'reality-oriented' cognition should therefore be a redundancy.

But cognition can also serve non-reality oriented ends, such as a defence-mechanism that is oriented to protecting ourselves psychologically from the negative experience of learning some truth of reality that is painful to accept. Moreover, engaging in reality-oriented cognition is never a given and can require a lot of effort and commitment. Engaging in reality-oriented cognition therefore requires an *active* commitment to want to know what is true. The kind of state that we have in mind can therefore be best described as a kind of meta-cognitive process, that is, thinking about our thinking, where we explicitly take a perspective on ourselves as epistemic agents with the chosen goal of knowing what is true. Such a state is both reality-oriented by being motivated that our cognition serves the end of knowing what is true and that our cognitive means serve that end, and such a state is active in that living up to such a motivation requires actively pursuing what is true over time. Gorlin call this state 'Cognitive integrity' because it involves having integrity with regard to using our minds to know the truth (Gorlin & Schuur, 2019).

We also describe Cognitive integrity as a disposition to refer to a person's temporal disposition to engage in the mental state of Cognitive integrity when circumstances call for it, such as when we are presented with epistemic difficulties and when we may be tempted to engage in cognitive processes that serve non-cognitive ends. The active and reality-oriented aspects of Cognitive integrity are in contrast to two other ways of using our minds.

Cognitive integrity is both active and reality oriented cognition in contrast to passive and non-reality-oriented cognition. In order to grasp what we mean by passive cognition, it is important to understand how cognition can be automatised and habitual. Most of the time we are not actively engaged in a meta-cognitive process of deliberating about our actions, but rather, much of our cognising is engaged in habitually and has been automatised over time (Wu, 2015).

The fact that much of our cognition is automatic in this sense is crucially important for us to learn new things and to act. If all of our cognition were slow and deliberate, then we would never get around to acting. However, precisely because we can automate our cognitive processes through habituation, we can also become passive with respect to ascertaining whether or not our cognitive processes succeed in and continue to serve as a tool to achieve knowledge of the truth. 'Passive cognition', in contrast to 'automated cognition', are instances where we default on the choice to ascertain whether or not our automated cognition continues to serve reality-oriented goals. This also contrasts with Cognitive integrity, where we are disposed to reassess whether our automated cognition serves reality-oriented goals, such as apprehending what is true.

Then there is a third way we can use our minds that we call 'pretence cognition', which is active with regards to the *means* of cognition but passive with regards to the end of cognition (Gorlin & Schuur, 2019). The end of being reality-oriented is not 'baked' into the nature of cognition, and cognition can be used for many non-reality oriented ends, such as when cognition serves the end of a psychological defence. For example, we may experience guilt for not living up to some of our goals. Subsequently, we can alleviate such guilt by giving ourselves excuses for why we could not achieve these goals, even though the reason that we did not achieve them is because of a fault of our own (such as lack of commitment, planning, or execution).

We call this use of our mind 'pretence cognition' because we are actively using the means of cognition to serve a non-reality oriented goal that we choose to be passive about, i.e., we are in a sense pretending to ourselves that we are engaged in reality-oriented cognition when we are in fact not engaged in reality-oriented cognition. Cognitive pretence is possible precisely because being aware of the ends of cognition itself requires an active choice, i.e., we need to choose to become aware what ends our cognition serves, since such ends are not at all obvious. This is true whether we are engaged in reality versus non-reality oriented cognition. All that being passive about the ends of our cognition requires is that we choose not to become aware of the ends of our cognition. It does not require any special kind of suppression of thought, it simply requires us to default in our thinking. Cognitive pretence is also made possible by the fact that the subjective experience of engaging in the means of cognition gives us some experiential basis to tell ourselves that we are engaged in reality oriented cognition. But such pretence is also often associated with subtle experiences of discomfort that what one is doing is not entirely honest. I will turn now to how Cognitive integrity provides a framework for solving the puzzle of addiction.

—3.3 Cognitive integrity as a framework for solving the puzzle of addiction

What then is the role of Cognitive integrity in the development and maintenance of knowledge of what is in our best interest? Whatever the content of what is in our best interest is, it requires not merely the discovery of facts about ourselves, such as our preferences, but also the development of our interests and other things we value. Knowledge of what is in our best interest is therefore not static but dynamic; we are required to grow it as we grow as persons and develop our interests over time. Once one gains knowledge of what is in one's best interest, keeping it is not guaranteed because you and the world continue to change. To the extent that knowledge of our best interest requires a willingness to gain such knowledge, the virtue of Cognitive integrity is indispensable for achieving, maintaining, and developing such knowledge of one's best interest.

But simply possessing knowledge of one's best interest in a theoretical sense does not guarantee that one will act on it. This is because such knowledge is not always 'front and centre' in our minds, especially since much of our thinking is habituated. Acting on knowledge of what is in our best interest requires us to disengage in habitual thinking and bring knowledge of what is in our best interest to our conscious awareness. Only then can we act on it. It is therefore possible to act contrary to our best interest despite having knowledge of it and how to achieve it because acting on such knowledge in its full context requires Cognitive integrity. This solution points to the fact that choice itself is involved in maintaining one's 'context of knowledge' when acting. Such a choice needs to be made every time one decides to act intentionally. There is no point where any choice guarantees that one will not maintain such a context of knowledge in the future. Gaining knowledge of what is in one's best interest is therefore never sufficient to act on such knowledge.

We can now see how either 'passive cognition' or 'cognitive pretence' can explain how we can fail to maintain the context of knowledge of what is in our best interest. In passive cognition we default on our choice to raise the level of awareness that is required to maintain our context of knowledge of what is in our best interest. For example, an addict may have recently gained knowledge that their desire to drink is triggered by memories of trauma, and that alternatives to drinking provide a better way of coping with such memories. But if the addict, when experiencing such memories again, defaults to his normal way of responding, especially if the alternative coping mechanisms require him to acknowledge those painful memories, then he will return to drinking to cope despite having knowledge that this is contrary to his best interests.

Through cognitive pretence we can drop the full context of the knowledge of what is in our best interest by subverting our cognition to non-reality oriented goals, such as denying to ourselves that we are acting contrary to our best interest. For instance, in the last example an addict could lie to themselves that facing those traumatic memories is not as bad as the negative consequences of continuing using drugs. Especially if an addict has past evidence that they always regret drinking after the fact and that they have evidence that alternatives to drinking are better at dealing with their negative emotions and at preventing their negative consequences, cognitive pretence can explain why an addict nevertheless continues drinking. This is possible not only because building a better life is hard in the face of past hardship, but also because acting on knowledge of one's best interest requires one to bring such knowledge to one's full awareness.

Cognitive integrity as an epistemic virtue provides a framework for solving the puzzle of addiction that does not deny that addicts have some knowledge of what is in their best interest (or knowledge of how to act on it). Neither does it deny that addicts don't have a strong enough will to act on such knowledge, as the Orthodox view does. Rather, from this framework, acting in your best interest requires one to maintain knowledge of what is in one's best interest (which requires choosing to be aware of it in action). The explanation for why addicts, or anyone, can act contrary to their best interest despite having knowledge of it and how to achieve it is twofold.

First, addicts can engage in passive cognition by being passive about the means of their cognition, thereby making themselves vulnerable to returning to drug use out of habit. Second, addicts can engage in pretence cognition by being passive about the end of cognition by actively subverting the means of cognition to non-cognitive ends, thereby justifying their continued drug use to themselves despite having knowledge that doing so is contrary to their best interest. Both forms of cognition are possible because wanting to know what is true itself requires choice.

The similarity between Cognitive integrity as a framework and Heyman's global choice framework is that both recognise the importance of having a long term perspective on one's best interest. Indeed, what makes a long term perspective on one's actions possible is the willingness to gain knowledge from such a perspective. This is shown by Pickard's work that knowledge of the causal relations between our actions and their consequences can only be gained if human beings have the capacity to make long term causal inferences. But Heyman takes the global choice framework as something that makes knowledge of one's best interest readily available and is sufficient for acting in your best interest. Again, the fact that we so often do not act in our best interest, despite being able to frame our actions from a long term perspective, is evidence for Heyman that we must all have some kind of 'natural bias' to the local choice framework. But from the perspective of the Cognitive integrity framework, the global choice framework is not sufficient for knowing what is in our best interest, nor is operating from it a given. Rather, the global choice framework is necessary but not sufficient for gaining knowledge of what is in one's best interest, as engaging the world with such a long term perspective itself requires choice. In other words, it is the local choice framework and being driven by habitual rather than self-aware thinking that is our default operating mode, whereas the global choice framework is a perspective that is both developed and must be chosen every time one wants to operate from it.

Where Heyman goes wrong is in assuming that the global choice framework ought to be sufficient for knowing what is in one's best interest and that acting from it is a given. But since our best interest is complex and requires both knowledge of what is contrary to it and what is for it, merely taking a long term perspective on one's self does not guarantee such knowledge. It is also because Heyman takes such a framework to be sufficient for acting in our best interest that he appeals to a 'natural bias' to the local choice framework. My solution posits that it is operating from the local choice framework that is our default because knowledge of and acting in our best interest is not a given because maintaining and developing Cognitive integrity is an achievement. We can thus fail to act in our best interest despite having knowledge of what is in our best interest and how to achieve it, as acting on such knowledge in context itself requires choice. This solution to the puzzle of addiction therefore builds on Heyman's work without inheriting its problems.

**Conclusion**

Heyman's work on addiction is important not only because it seeks to explain addictive behaviour, but because it seeks to explain a more general kind of problem that is relevant within the context of medicine: voluntary harmful behaviour. Explaining this kind of behaviour is important because we want to help patients recover. But we also want to avoid undermining or over-emphasising the voluntary control that patients have over the behaviours in question.

Heyman has argued that, in order to make room for voluntary explanations of addiction, we have to question our fundamental assumptions regarding the nature of voluntary behaviour. In particular, Heyman argues that the assumption that voluntary behaviour is necessarily rational with respect to acting in our best interest necessarily leads to the conclusion that irrational behaviour with respect to our best interest, i.e., self-destructive behaviour, is involuntary. I have called this view homo economicus. Heyman argues that because addictive behaviour responds to incentives, addictive behaviour is therefore voluntary. Both the Orthodox view of addiction and the view of homo economicus therefore appear to be false, according to Heyman.

But this still leaves unexplained the puzzle of addiction, namely: why do addicts keep using drugs when doing so is detrimental to their best interest. Heyman's solution to the puzzle of addiction is twofold. First, Heyman points out that how we temporally frame our actions changes how we assess their outcomes, i.e., focusing on short term benefits blinds us from the long term costs of using drugs. But this distinction between the local and global choice frameworks raises the puzzle of why addicts, or everyone, does not always operate from the global choice framework. Second, Heyman argues that we have a 'natural bias' to frame our actions from the local choice framework, and that if it weren't for this 'natural bias' that we would all be operating from the global choice framework. I have called this view homo irrationalis, that view that we have a natural bias to frame our actions from the local choice framework. While Heyman's model of addiction and theory of voluntary behaviour does provide a solution to the puzzle of addiction, it also faces serious problems. Yet, it also points to real progress that can help explain harmful voluntary behaviour, such as the local versus global choice framework distinction.

The purpose of this paper has been to both criticise and build on Heyman's solution to explain how harmful behaviour can be voluntary. My main criticism of Heyman's solution is that the appeal to a 'natural bias', and homo irrationalis, faces many of the same problems that face the weak Orthodox view of addiction. I have argued that there are two versions of the Orthodox view of addiction, a strong and weak version. The strong version holds that addictive behaviour is involuntary in the way that Heyman defines it. This version is often defended by making various distinctions that claim compatibility with the evidence that shows that addictive behaviour does change in response to incentives. For instance, arguing that addiction has an acute and a chronic phase or stage. The weak version of the Orthodox view rejects Heyman's definition of involuntary behaviour and argues that addictive behaviour is involuntary in same broader sense. The problem with this position is that it is not clear what is meant by involuntary behaviour on the weak Orthodox view and therefore that the view loses some of its explanatory power to solve the puzzle of addiction. I have argued that Heyman's appeal to a 'natural bias' faces the same problem.

Heyman's 'natural bias' plays a crucial part in his solution to the puzzle of addiction, yet he never defines it. Moreover, it is unclear from Heyman's model what the relationship is between the addict and the local and global choice frameworks. Can the addict's behaviour only be changed in response to incentives, or does the addict have some control over how they frame their actions? That is, can the addict choose to frame his actions differently such that they are subject to incentives in different ways? If not, then it is hard to see how Heyman's model is any different from the weak Orthodox view. If so, we need an account of how this capacity is possible.

Heyman argues that we should reject the view that voluntary behaviour is necessarily rational with respect to acting in our best interest and replace it with the view that voluntary behaviour is biased towards irrationality with respect to acting in our best interest. But this view may actually be the assumption underlying the Orthodox view. What these all these views have in common is the assumption that something like the global choice framework ought to be sufficient for knowing what is in our best interest and that such knowledge ought to be sufficient for acting in our best interest. Because we so often act contrary to our best interest despite having the capacity to know it, Heyman infers that we must have some 'natural bias' to act against it.

But the real assumption that should be challenged here is that the global choice framework is sufficient for knowing what is in our best interest and that knowledge of what is in our best interest ought to be sufficient for acting in our best interest. I have drawn on the work of Hanna Pickard on the role of denial in addiction to show that knowledge of what is in an addict's best interest does not materialise simply by the addict taking a long term perspective on themselves and their actions. Rather, taking such a perspective makes such knowledge possible but not inevitable. Moreover, merely having knowledge that continued drugs us is contrary to one's best interest is not sufficient to stop using, since one also requires positive knowledge of what is in one's best interest and how to achieve it, such as by developing drug-free alternative behaviours and believing that a better life is possible. Pickard also shows that denial can explain why addicts keep using drugs despite knowledge of the severe negative consequences. But while denial can explain why addicts keep using drugs, it doesn't explain why addicts return to using drugs once they have accepted that continued drug use is bad; especially once they have gained knowledge of what is in their positive best interest (such as drug-free alternative behaviours).

I have argued that the epistemic virtue of Cognitive integrity, as developed by Gorlin, provides a framework for explaining how we can act contrary to our best interest despite having knowledge of it and how to achieve it (Gorlin & Schuur, 2019). Cognitive integrity refers to the mental state and virtue of one's willingness to gain knowledge. To the extent that gaining and maintaining knowledge of one's best interest is a major achievement it requires Cognitive integrity. But even when one has such knowledge, possessing it alone is not sufficient for acting on it. This is because knowledge of what is in one's best interest must always be brought back to one's level of awareness in order to be actionable, which itself is a virtue that must be developed. One can therefore fail to act in one's best interest despite having knowledge of one's best interest because acting on such knowledge requires of us to bring such knowledge to one's conscious awareness. This needs to be done each time we need to act. This solution does not rely on a 'natural bias' but instead acknowledges that gaining and acting in one's best interest is always an achievement.

## References


Abramowitz J. S. 2006. The psychological treatment of obsessive-compulsive disorder. The Canadian Journal of Psychiatry. 51,7: p. 407-16.

Alston, W. 1988. The deontological conception of epistemic justification. Philosophical Perspectives, 2, p. 257–299.

American Psychiatric Association (APA). 2013. Diagnostic and Statistical Manual of Mental Disorders, 5th *edition*, Washington, DC: American Psychiatric Association.

Audi, R. 2001. Doxastic voluntarism and the ethics of belief. In M. Steup (Ed.), *Knowledge, Truth and Duty*. New York: Oxford University Press. p. 93-114.

Bargh, J. A. 1994. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. S. Jr (Eds.), *Handbook of social cognition, Vol. 1: Basic processes; Vol. 2: Applications (2nd ed.)* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. p. 1-40.

Bortolotti, L. 2022. Are delusions pathological beliefs?. *Asian Journal of Philosophy* 1,31.

Chavigny, K. A. 2013. “An Army of Reformed Drunkards and Clergymen”: The Medicalization of Habitual Drunkenness, 1857–1910. *Journal of the History of Medicine and Allied Sciences* 69:3, p. 383-425.

Crothers, T. D. 1891. Are Inebriates Curable? *JAMA.* 17,24: p. 923-927.

Evans, J. S., & Stanovich, K. E. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives in Psychological Sciences, 8:*3, p. 223-241.

Ferentzy, P. 2001. From sin to disease: differences and similarities between past and current conceptions of chronic drunkenness. Contemporary Drug Problems 28, p. 363-390.

Fingarette, Il. 1988. Heavy drinking: The myth of alcoholism as a disease. University of California Press.

Foley, R. 1999. Voluntarism. In R. Audi (Ed.), The Cambridge dictionary of philosophy 2nd ed. New York: Cambridge University Press. p. 964.

Frank, L. E. & Nagel, S. K. 2017. Addiction and Moralization: the Role of the Underlying Model of Addiction. *Neuroethics* 10: p. 129–139.

Gorlin, E. I., & Schuur R. 2019. Nurturing Our Better Nature: A Proposal for Cognitive Integrity as a Foundation for Autonomous Living. Behavior Genetics. p. 154–167.

Heather, N. 2013. Is alcohol addiction usefully called a disease? *Philosophy, Psychology, & Psychiatry* 20:4, p. 321-324.

Heil, J. 1983. Doxastic agency. *Philosophical Studies, 43,* p. 355-364.

Hershon, H. 1974. Alcoholism and the Concept of Disease. *British Journal of addiction* 69, p. 123-131.

Heyman, G. M. 1996. Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19, p. 561-610.

Heyman, G. M. 2009. Addiction: A Disorder of Choice. Cambridge, MA, Harvard University Press.

Heyman, G. M. 2013. Addiction and choice: theory and new data. *Frontiers in psychiatry: Perspective Article,* 4:31, p. 1-5.

Heyman, G. M. 2017. Do addicts have free will? An empirical approach to a vexing question. Addictive Behaviors Reports 5, p. 85–93.

Hieronymi, P. 2006. Controlling attitudes. *Pacific Philosophical Quarterly, 87*, p. 45–74.

Jellinek, E. M. 1946. Phases in the drinking history of alcoholics. *Quarterly Journal Studies on Alcohol* 7, p. 1-88.

Jellinek, E. M. 1952. Phases of alcohol addiction. *Quarterly Journal of Studies on Alcohol* 13, p. 673-84.

Jellinek, E. M. 1960a. Alcoholism: A genus and some of its species. *Canadian Medical Association Journal.* 83: p. 1341–1345.

Jellinek, E. M. 1960b. The disease concept of alcoholism. Highland Park, NJ; Hillhouse.

Johnstone, G. 1996. From vice to disease: The concepts of dipsomania and inebriety, 1860-1908. Social Legal Studies, 5.1, p. 37-56.

Kennett, J. 2013. Addiction, Choice, and Disease: How Voluntary Is Voluntary Action in Addiction? In Neuroscience and Legal Responsibility, edited by Nicole A. Vincent, OUP. p. 258-278.

Leshner, A. I. 1997. Addiction Is a Brain Disease, and It Matters *Science* 278, p. 45-47.

Leshner, A.I. 2001. Addiction is a brain disease. Issues in Science and Technology Online, 17,3.

Levine, H. G. 1978. The discovery of addiction: Changing conceptions of habitual drunkenness in America. *Journal of Studies in Alcoholism* 39: p. 143-174.

Lewis, D. C. 1991. Comparison of Alcoholism and Other Medical Diseases: An Internist's View. *Psychiatric Annals* 21,5; p. 256-265.

Lewis, M. 2018. Brain Change in Addiction as Learning, Not Disease. *The New England Journal of Medicine* 379;16, p. 1551-1560.

Lynch, K. 2014. Self-deception and shifts of attention, Philosophical Explorations, 17:1, p. 63-75.

Maltzman, I. 1991. Is Alcoholism a Disease? *A Critical Review of a Controversy. Integrative Physiological and Behavioral Science*, 26,3, p. 200-210.

Mele, A. R. 1986. Self-Deception and “Akrasia”. Behaviorism, 14,2 p. 183-192

Mele, A. R. 1997. Real Self-Deception. *Behavioral and Brain Sciences*, *20*, p. 91–102.

Miller, N. S. and Chappel, J. N. 1991. History of the Disease Concept. *Psychiatric Annals* 21,4; p. 196-205.

Miller, W. R., & Rollnick, S. 1991. *Preparing people to change addictive behavior*. New York, NY: Guilford Press.

Miller, W. R., Westerberg, V. S., Harris, R. J., & Tonigan, J. S. 1996. What predicts relapse? Prospective testing of antecedent models. *Addiction*, *91*, p. 155-171.

Olatunji B. O, Davis M. L, Powers M. B, Smits J. A. 2012 Cognitive-behavioral therapy for obsessive-compulsive disorder: a meta-analysis of treatment outcome and moderators. Journal of Psychiatry Research. 47,1: p. 33-41.

Pickard, H. 2012. The Purpose in Chronic Addiction. *AJOB Neuroscience*, 3:2, p. 40–49.

Pickard, H. 2013. Psychopathology and the ability to do otherwise. *Philosophy and Phenomenological Research,* p. 135-163.

Pickard, H. 2016. Denial in Addiction. *Mind & Language*, 31:3, p. 277–299.

Pickard, H. 2018. The Puzzle of Addiction. In The Routledge Handbook of Philosophy and Science of Addiction. Eds. H. Pickard and Serge H. Ahmed. p. 9-22.

Pickard, H. & Ahmed, S. H. 2016. How do you know you have a drug problem? The role of knowledge of negative consequences in explaining drug choice in humans and rats. In 'Addiction and Choice: Rethinking the relationship' Eds. Heather, N. And Segal, G. Oxford University Press. p. 29–48.

Prelec, D. & Herrnstein, R. J. 1991. Preferences or principles: alternative guidelines for choice. In 'Strategy and Choice'. Ed. R. Zeckhauser. Cambridge, MA: MIT Press. p. 319-340.

Racine, E., Bell, E., Zizzo, N., & Green, C. 2015. Public Discourse on the Biology of Alcohol Addiction: Implications for Stigma, Self-Control, Essentialism, and Coercive Policies in Pregnancy. *Neuroethics* 8: p. 177–186.

Rush, B. 1805. An Inquiry into the Effects of Ardent Spirits upon the Human Body and Mind, 4th edition. Philadelphia: Printed for Thomas Dobson, 1805.

Rush, B. 1812. Medical Inquiries and Observations upon the Diseases of the Mind. Philadelphia: Kimber & Richardson.

Salmieri, G., & Bayer, B. 2014. How we choose our beliefs. *Philosophia, 42,* p. 41-53.

Satel, S., & Lilienfeld, S. O. 2014. Addiction and the brain-disease fallacy. *Frontiers in Psychiatry*: Review Article, 4:141, p. 1-11.

Segal, G. M. A. 2013. Alcoholism, disease and insanity. *Philosophy, Psychiatry, & Psychology* 20,4: p. 297–315.

Sharples (Trans.), R. 2007. *Alexander of Aphrodisias on Fate*. London: Duckworth Publishers.

Spitzer, R. L. & Endicott, J. 1978. Medical and mental disorder: Proposed definition and criteria. In: Spitzer, R. L. & Klein, D. F., eds. Critical Issues in Psychiatric Diagnosis. New York, NY: Raven Press; p. 15-39.

Steup, M. 2011. Belief, Voluntariness and Intentionality. *Dialectica,* 65:4, p. 537-559.

Stichter, M. 2013. Virtues as skills in virtue epistemology. *Journal of Philosophical Research, 38,* p. 333-348.

Street, M. D., Douglas, S. C., Geiger, S. W., & Martinko, M. J. 2001. The impact of cognitive expenditure on the ethical decision-making process: The cognitive elaboration model. *Organizational Behavior and Human Decision Processes, 86,* p. 256-277.

Todd, J. E. 1882. Drunkenness a vice,—not a disease; a paper by Rev. John E. Todd. Hartford; Case, Lockwood & Brainard.

Vaillant, G. E. & Milofsky, E. S. 1982a. The etiology of alcoholism. *American Psychologist* 37, p. 494-503.

Vaillant, G. E. & Milofsky, E. S. 1982b. Natural history of male alcoholism: 4. Paths to recovery. *Archives of* General Psychiatry 39, p. 127-33.

Vonasch, A. J., Clark, C. J., Lau, S., Vohs, K. D., & Baumeister, R. F. 2017. Ordinary people associate addiction with loss of free will. *Addictive Behaviors Reports,* 17:5, p. 56-66.

Warner, J. 1994. Resolv'd to drink no more: Addiction as a preindustrial construct. Journal of Studies on Alcohol 55, p. 685-691.

Williams, B. 1973. Deciding to believe. In B. Williams ed., Problems of the self. Cambridge: Cambridge University Press. p. 136–151.

Wu, W. 2015. Experts and Deviants: The Story of Agentive Control. *Philosophy and Phenomenological Research,* 93:1, p. 101-126.

**Rejecting the Disease Versus Moral-Problem-in-Living Distinction to Constrain Medicine**

Modern medicine has provided tremendous value to humanity by alleviating suffering and improving our lives. But modern medicine has also raised concerned about how medicine relates to other problems in life and the medicalisation of such problems (Szasz, 1961). One such concern is that problems of a voluntary and moral nature are 'medicalised' and that medical problems are 'moralised'. For example, it has both been argued that addiction is moralised despite being a 'disease', and that addiction is medicalised despite being a 'moral-problem-in-living' (Leshner, 1997; Szasz, 1972). The shared view is that the problems that medicine and morality deal with are fundamentally different in kind and that therefore there should be a clear separation between these two domains. Indeed, throughout history there have been concerns about the misapplication of medical and moral explanations, and of medical and so-called 'moral treatments' to each other. Typical examples are over-attributing agency to patients and blaming them for their illness, or under-attributing agency to people and disempowering them. There are therefore good reasons to prevent a misapplication of narrowly medical versus moral approaches.

One solution to this problem has been to separate the domains of medicine and morality by clearly demarcating medical problems from moral-problems-in-living, where moral-problems-in-living are problems of the self that are subject to voluntary guidance and diseases are problems of an involuntary nature and therefore not subject to such guidance (Szasz, 1961). Such a distinction seeks to prevent the misapplication of strictly medical and moral explanations and treatments by removing or reducing voluntary explanations within medicine. This distinction can be seen in medicine when it was argued that addiction is a disease and not a moral-problem-in-living *because* addictive behaviour is involuntary (Miller and Chappel, 1991 p. 196-197). To prevent the moralisation of medicine, medicine is defined as primarily concerned with diseases as involuntary and non-moral and therefore not concerned with voluntary moral-problems-in-living. I call the use of this distinction to constrain medicine the *medicine versus morality distinction*.

But using this 'disease versus moral-problem-in-living' distinction to delimit the scope of medicine also has certain negative consequences. Viewing diseases as involuntary creates incentives towards involuntary explanations of diseases in medicine. For example, if addiction's medical status is in part justified on the view that addictive behaviour is involuntary, then addiction's medical status is threatened if addictive behaviour turns out to be voluntary. This reveals a potential dilemma between explaining a condition and its medical status. For example, viewing addiction as an involuntary disease incentivises funding into research and treatments that conform to such views. Some have called this a 'Faustian bargain' because calling addiction an involuntary disease may increase funding, but direct that funding towards less effective treatments (Satel and Lillienfeld, 2014, p. 4). Viewing diseases as involuntary has also created a simplistic framework for dealing with issues of stigma and blame because fighting stigma and blame would then require us to undermine the role that agency might play, since attributing agency is tied to moralisation, blame, and stigma (Racine et al., 2015). Viewing diseases as involuntary as a means to prevent 'moralising' diseases therefore has non-trivial costs, such as incentivising involuntary explanations and disincentivising voluntary explanations in medicine.

But this ‘hard’ disease versus moral-problem-in-living distinction has already partially been rejected because it is based on an overly narrow conception of disease as defined by an involuntary aetiology, i.e., causes of an involuntary nature. Indeed, since the 1970’s it has been argued that diseases should not be defined primarily by their causes but instead by their negative consequences (Kendell, 1975). Much of the debate about the concept of disease has focused on what these negative consequences should be and how to distinguish them from non-medical negative consequences, such as a biological dysfunction (Boorse, 1975; Wakefield, 1992).

Moral-problems-in-living can therefore be diseases if they qualify as a disease generally, i.e., they involve a medically relevant negative consequence. But allowing for diseases with a voluntary or moral nature does not imply a rejection of the strategy of separating medicine and morality to prevent misapplications between them. Rather, the hard version of the disease versus moral-problems-in-living distinction has been replaced with a soft version: that diseases can have moral components if they also have ‘indispensable non-moral components’ (Sadler, 2005). This soft distinction seeks to combine the utility of voluntary and moral explanations within medicine while still trying to separate between medicine and morality to prevent some misapplications.

The purpose of this paper is to argue that we should reject the soft disease versus moral-problem-of-living distinction as a strategy of preventing misapplication of medicine and morality to each other by separating the two domains. Rather, I argue that we should allow for more overlap between medicine and morality and prevent any misapplications by better distinguishing between non-moral versus moral-problems-in-living within medicine. That is, I argue that we should only use this distinction as a means to distinguish between different kinds of explanation within medicine, not to use such a distinction to separate medicine and morality as such. The purpose of this paper is not, however, to give a full account of what the disease versus moral-problem-in-living distinction as explanatory should look like. My purpose is to argue that using the soft version of the distinction to separate medicine and morality disincentivizes developing any such account. I define moral-problems-in-living as problems facing those aspects of the self that are open to voluntary control that require normative guidance with respect to ultimate ends, such as some conception of a good human life. This broad definition allows me to be clear about what the distinction refers to generally without having to commit to its substantive development. I will clarify the different uses of the term ‘moral’ in the literature, including my own, in the next section.

This paper is divided into three parts. In part one I introduce the disease versus moral-problem-in-living distinction as an explanatory distinction within medicine and then as a distinction used to separate medicine and morality. I argue that using this distinction to separate medicine and morality creates a dilemma between voluntary/moral explanations and the medical status of certain conditions, like addiction. In part two I distinguish between the hard and soft version of the disease versus moral-problem-in-living distinction and explain that the hard version of the distinction has been rejected because it is based on a too narrow conception of disease. I explain how the soft version of the distinction allows for some moral explanation in medicine while still seeking to separate medicine and morality. In part three, I argue that the soft version of the distinction to separate medicine and morality should be rejected and that we should prevent misapplication by distinguishing between moral and non-moral explanations within medicine.

## —1.0 The disease versus moral-problems-in-living distinction

The disease versus moral-problem-in-living distinction is rarely discussed on its own terms and mostly comes up in discussions of particular conditions or in criticisms of psychiatry (Charland, 2004; Szasz, 1961). The distinction can refer to many issues and is ill-defined. While I will clarify the disease versus moral-problems-in-living distinction, this lack of clarity is not only due to neglect. How to distinguish between diseases and moral-problems-in-living is not always clear precisely because the meaning of these terms are contested. The legitimacy of the distinction is also often taken for granted and its motivations left implicit. One of my goals in this paper is to make explicit those motivations to help inform us on how to think about this distinction. But first I will dismiss two ways that the disease versus moral-problems-in-living distinction has been used.

Sometimes the disease versus moral-problems-in-living distinction is used to distinguish between 'diseases' as legitimate medical disorders and 'moral judgments' of phenomena that are either (1) considered legitimate medical disorders but that should not be regarded as legitimate medical disorders or (2) phenomena that are not considered legitimate medical disorders but that should be regarded as legitimate medical disorders. A 'moral' problem under this distinction captures a sociological fact that some phenomenon, usually a behaviour, is evaluated as 'bad' by society as large (Agich, 1994). Drapetomania, the 'medical' condition of slaves running away in 19th century America, is an example of behaviour that was regarded as a disease to be controlled by a society that judged such behaviours as being 'bad' (Szasz, 1961). The other misjudgement concerns legitimate medical conditions that are *denied* medical status because patients are regarded as 'bad', such as regarding the severely mentally ill as being 'bad' rather than having medical conditions. The problem with this use of the distinction is that the term 'moral' is simply used to refer to norm and values generically. It is a legitimate issue not to medicalise a condition or deny medical status merely because it is disliked for other reasons. But this use of the term moral obfuscates the distinct issues under the header of what I call 'moral-problems-in-living'.

Another way that the disease versus moral-problems-in-living distinction is used is to refer to the issue of determining whether a person can be regarded as criminally responsible for their actions, or is found to be 'not guilty' by 'reason of insanity'. This is also referred to as the 'mad versus bad distinction' (Sadler, 2008). Here 'moral' is used to refer primarily to criminal and illegal actions, behaviours, and activities. Being 'moralised' here is therefore not capturing a sociological fact that some behaviour is merely disliked by society and which might very well be 'healthy' behaviour. Rather, the concept of 'bad' is intended to capture distinct states of intention, control, and knowledge of good and bad that bear on the substantive question of criminal responsibility.

While the first use of the disease versus moral-problem-in-living distinction is too broad, this second use of the distinction is much too narrow. The question of whether a person is criminally responsible for their actions by reason of insanity is a legitimate issue, but this question concerns but a small area in our 'moral universe'. The question is independent of whether medicine should be concerned with non-criminal moral problems. This use of the disease versus moral-problem-in-living distinction also has a very narrow conception of 'moral' as referring to criminal behaviour. This use of 'moral' misses the wider phenomena of 'moral-problems-in-living'.

What do I mean by 'moral-problems-in-living', then? Morality concerns the evaluation of actions and persons in some normative sense as being right or wrong, good or bad. But morality is not just concerned with evaluation *as such*, in a generic sense. Morality as a discipline intends to provide evaluations for the purpose of giving guidance for action: of giving prescriptions for what we ought to do. Morality is therefore concerned with providing guidance towards certain ends. What appears to distinguish these ends from other ends is that they are in some sense ultimate ends, such as what kinds of persons we want to be or what kind of society we want to live in. Moral philosophies can differ on what ultimate ends we should strive for, whether it be the greatest happiness of the greatest number, living up to certain moral principles for their own sake, or cultivating a particular kind of character and striving towards our own happiness. While moral philosophies can differ on the ends morality should serve, they all propose such ends.

But the prescriptive nature of ultimate ends in morality presumes that we are agents that are free to act on the guidance that morality provide us. 'Should' and 'ought' implies 'could' and 'can'. That is, morality concerns issues that we have some kind of control over as agents. This is why agential concepts such as free will, choice, volition, action, autonomy, and agency are so central to morality, because it is only to free agents that morality can provide normative guidance.

Some moral philosophies will be concerned with giving guidance for a particular phenomenon that we have control over, such as our actions and whether our actions are right or wrong in themselves. But this narrow focus of morality on right and wrong action was challenged by the middle of the 20th century for being impoverished of the actual scope of guidance that morality can and should provide to agents (Anscombe, 1958). The criticism was that morality need not only be concerned with right or wrong actions but that it can provide guidance on other aspects of the self, such as our character, and other goals, such as our own flourishing. So not only are actions the subject matter of moral guidance but so are other aspects of the self that we can guide and develop through voluntary means, such as our emotions, beliefs, and how we choose to use our minds. This criticism of an overly narrow conception of morality has been noted in the debate regarding the medical status of Cluster B Personality Disorders[56].

One author has noted that an overly narrow conception of morality as focused only on right or wrong action necessarily generates a medicine versus morality distinction[57]. I define moral-problems-of-living in this paper as problems facing those aspects of the self that are open to voluntary control that require normative guidance. Moral explanations are therefore those explanations that explain phenomena via aspects of the self that are subject to moral guidance, such as how lacking knowledge of the good or virtues can explain some harmful behaviours. I will now turn to how the disease versus moral-problem-in-living distinction has been used as a distinction to explain certain conditions and to separate the domains of medicine and morality.

---

[56] "Aristotle's virtue theory and his notion of ethics…is broader than entrenched legalistic ideas about 'morality.' Ethics is less about what we should do and more about what kind of persons we should be" (Zachar and Potter, 2010 p. 108).

[57] "Incoherence arises because of the conflict between modern conceptions both of morality and of health and illness and the necessary use of moral terms in defining what makes some mental disorders. Modern moral philosophy and modern conceptions of health and illness imply that health is a *non-moral* good, and so that illness is a "disorder" in a non-moral sense. *Ancient ethics, on the other hand, allows states as well as actions to be morally evaluated, and so for the notion of a moral order or disorder which is not consciously chosen*" (Matthews, 1999 p. 299 emphasis added).

—1.1 The disease versus moral-problem-in-living distinction as explanatory

The disease versus moral-problem-in-living distinction has been used to refer to a distinction between diseases as involuntary conditions and therefore non-moral in nature and as conditions that are voluntary and therefore moral in nature (Lewis, 1991). This distinction can function both to explain conditions within medicine and/or separate medicine from morality as such. But these two functions can come apart and even come into conflict with each other. While the exact relation between volition and moral problems is not always clear, I will present a basic account that is intended to capture how this distinction tends to function in the literature. I will first present the explanatory function of the distinction before presenting its demarcating function.

The concept of disease has historically been used in two ways. First, in a broad sense to refer to a collection of symptoms. Second, in a narrow sense to refer to the causes of those symptoms (Lewis, 1991). To claim that someone has a disease in the broad sense is to refer to their symptoms, and therefore this concept of disease does not function as an explanation. Rather, in this broad sense, disease refers to the phenomenological experience of 'not being at ease'. After the germ theory of disease, however, the concept of disease was beginning to also be used to refer to the causes of diseases and thereby served as an explanation of the symptoms under scrutiny. To claim that someone had a disease could therefore refer to their symptoms and/or refer to the causes which explain their symptoms. Those causes of symptoms often referred to discrete entities like germs or lesions (Kendell, 1975; Pies, 1979). The concept of disease was therefore often used to imply some kind of involuntary process that the patient had no direct control over. There are many examples of people referring to diseases as being involuntary[58] and patients as victims[59]. The concept of disease as involving involuntariness is therefore also intended to have explanatory power. The best example of using the concept of disease to causally explain a condition is addiction[60]. Addiction presents a puzzle as to why addicts keep using drugs despite their severe negative consequences. Addiction being an involuntary disease solves this puzzle by arguing that addicts keep using drugs despite their severe negative consequences because they have lost control over their behaviour (Leshner, 1997). But the extent to which addiction as an involuntary disease is involuntary is not always clear. Some have argued that addiction is not a disease because addictive behaviour is not involuntary, but voluntary. Others, however, argue that critics of the disease model of addiction have misread the disease model of addiction (Heyman, 2009; Segal 2013; Kennet, 2015). I will not get into those disputes, but suffice it to say that the concept of disease as involuntary has been used to explain certain conditions, like addiction, even though it is not always clear what these terms mean.

---

[58] "…diseases — that is, phenomena independent of human motivation and will" (Szasz, 1961 p. 275). "If…the etiology and/or course of the condition is primarily beyond the control or intention of the individual, it is more acceptable to call that condition a disease" (Lewis, 1991 p. 256-258). "…if a key feature of a disease is that the symptoms are involuntary, then we need to know how to distinguish between voluntary and involuntary behavior" (Heyman, 2009 p. 90).

[59] "Rarely overtly stated but clearly central to the concept of a disease is a victim state. As a victim, the afflicted has no control over the progression of the disease if left untreated" (Miller and Chappel, 1991 p. 197).

[60] "The loss of control, which can actually be inherited, is the sine qua non for alcoholism (and drug addiction) as qualifying for the disease state" (Miller and Chappel, 1991 p. 197).

Moral concepts, such as vice, sin, and habit, have also been used to describe and explain conditions under the purview of medicine. The clearest example of this is also, again, descriptions and explanations of addiction, or 'habitual drunkenness' (and later inebriety and dipsomania) (Warner, 1994). 17th century English clergymen called addiction a disease, habit, vice, and a sin, often all at the same time. Applying both medical and moral terms was not only done by the clergy but also by many later physicians, such as Benjamin Rush (Rush, 1812). Even in recent decades some have argued that certain condition under the purview of medicine have indispensable moral components, such as the Cluster B Personality Disorders, which are described in moral terms such as lacking empathy, not respecting the rights of others, and being manipulative (Charland, 2004). This raises the question, what does 'moral' mean in this context?

Moral concepts appear to have been applied to conditions under the purview of medicine in two ways: as descriptions and as explanations of certain conditions. Descriptions of a condition in moral terms did not imply anything about whether the condition itself was what I call a moral-problem-in-living, but only indicated that certain aspects of the self, or moral self, were affected by the condition in question. The term 'moral' in this broad sense is used to refer to those aspects of our psychology that are subject to development and guidance (Charland, 2008 p. 20).

One example of the use of term 'moral' in this descriptive sense in the 19th century was the condition of 'moral insanity'. 'Moral insanity' was regarded as a particular kind of insanity where the person in question was not able to know right from wrong or where such a person committed heinous and criminal acts despite knowing right from wrong. In either case, 'moral insanity' was regarded as a condition where the 'moral sense' or conscience was either missing or impaired. Then, as still is the case today, it was disputed whether moral insanity was due to some disease of the brain or whether it was due to some failure in proper moral development[61].

This notion of a 'moral condition' is therefore fully compatible with a condition being a disease. More importantly, in response to the question of whether 'moral insanity' is a vice or a disease, Benjamin Rush was clear that the answer to such a question should have no bearing on whether or not 'moral insanity' ought to be the object of social compassion or medical concern[62]. Indeed, Rush draws several similarities between diseases of the body and vices of our moral faculties, such as the idea that both can be caused by inaction and neglect of the self[63]. Rush points to such similarities to argue that problems in our moral faculties, whatever their causes may be, whether they are 'of necessity' or of 'agency', should be under the purview of medicine.

---

[61] "In all these cases of innate, preternatural moral depravity, there is probably an original defective organization in those parts of the body, which are occupied by the moral faculties of the mind. How far the persons whose diseases have been mentioned, should be considered as responsible to human or divine laws for their actions, and where the line should be drawn that divides free agency from necessity, and vice from disease, I am unable to determine" (Rush, 1812 p. 358).

[62] "In whatever manner this question [of the nature of moral insanity] may be settled, it will readily be admitted that such persons are, in a pre-eminent degree, objects of compassion, and that it is the business of medicine to aid both religion and law, in preventing and curing their moral alienation of mind. We are encouraged to undertake this enterprise of humanity, by the sameness of the laws which govern the body and the moral facilities of man" (Rush, 1812 p. 358).

[63] "Is debility the predisposing cause of disease in the body? So it is of vice in the mind. This debility in the mind consists in indolence, or a want of occupation… The near relation of debility and vice has been expressed by the schoolmen in the following words,… To do nothing is generally to do evil" (Rush, 1812 p. 358-359).

But the concept of 'moral' was also used as an explanation. In much the same way that the narrow concept of disease was used to explain the symptoms of disease, this narrow concept of 'moral' was used to explain the causes of a condition that was described in moral terms. Not all conditions that were *described* in moral terms were *explained* in moral terms, like moral insanity. Explanations in moral terms often involved some failure of development of those aspects of the self that could be directed to ultimate ends by one's actions, such as one's habits, emotions, or character. Importantly, such explanations did not necessarily imply blame on part of the person in question, since moral development was considered to be a product of moral education, which individuals were not always blamed for lacking. One could therefore fail to develop those aspects of the self through a variety of means, such as a lack of knowledge of the good, of motivation to gain such knowledge, or a lack of opportunity to act on such knowledge. The best example of this kind of explanation is again of moral explanations of addiction (Warner, 1994). Such explanations argued that addicts keep using drugs despite severe negative consequences because the more one used, the more one gave into one's temptations. This process subsequently weakens one's will for future resistance and thereby makes it harder to resist the more one used. While framed in moral terms such explanations are remarkably similar to behaviourists explanations of addiction, as in part spurred on by habituation (see Heyman, 2009).

Sometimes such moral terms were used interchangeably with medical terms, such as in the case of the 17th century English clergymen and Benjamin Rush. This was most likely because they would be using the concept of a disease in a descriptive sense while using moral concepts in an explanatory sense. At other times, the concept of disease and moral concepts were used in mutually exclusive ways to describe the same phenomena. For example, Reverend Todd argued that the concepts of disease and vice are mutually exclusive and should therefore be separated:

"… the impulse to drink either can or cannot be resisted, so far as the state of the physical system is concerned; if it can, the yielding is vicious; if it can not, the yielding is the result of disease. The same act can hardly be both physical and moral, both necessary and voluntary, both criminal and innocent, both the result of the working of physical laws and forces and the result of the exercise of volition, at the same instant" (1883 Todd, p. 2).

However, while Reverend Todd points to the role that volition plays in serving to distinguish disease from moral explanations, he also argues that this role is not essential because, from a certain perspective, even moral explanations can be partially involuntary explanations:

"Granting, for the sake of the argument, that in some extreme cases, or even in many or all cases if you choose, the will is helpless. *This is not necessarily or probably the effect of disease.* For it is in the nature of vice to become inveterate. A man may reject the gospel to such a point that it is impossible for him to repent; but it is not a disease with him. Any habit may become a despotic tyrant, an irrefragable chain; but habits are not diseases" (1883 Todd p. 9-10 emphasis added).

This last quote from Reverend Todd raises interesting questions of what is really meant by a vice being involuntary. It could simply mean that a vice is a habit that is formed over time and as such, once formed, cannot simply be unchosen 'at will', but will require repeated actions to undo. Or Reverend Todd could simply be referring to the doctrine of original sin, which itself is a quasi-medical explanation for why people so often engage in immoral behaviour, namely, a 'sick' soul. In either case, we will see in the next section that both those within and outside of medicine also have non-explanatory reasons to separate medical from non-medical explanations and approaches.

But for much of the 19th century, both strictly medical and moral explanations were used within medicine. Another good example of moral terms being used as descriptions and explanations within medicine is the area of affective psychopathology in the 19th century, especially in the works of Pinel on various affective disorders (Charland, 2008). Pinel's view of certain affective disorders is that they were disorders in the moral senses and the passions. In particular, Pinel viewed some instances of 'moral insanity' as cases where it was primarily an affective rather than a cognitive disorder, and that treatment required affective conditioning. Much of Pinel's work on affective psychopathology and his 'moral treatment' has been interpreted as being primarily psychological in nature, but not as being 'moral' in a narrower sense of the term. However, Charland has argued that Pinel did view both affective disorders and moral treatment as being 'moral' in a more narrow sense than simply psychological[64]. Charland, for example, points out that much of Pinel's explanations and treatments depend on an explicitly moral framework. That is, Pinel's framework was focused on explaining and intervening on affective disorders as problems in the formation of a moral self, where morality has to do with giving guidance to certain ends in developing one's character. In particular, Pinel focused on issues of benevolence, sympathy, and self-esteem as objects of affective conditioning in his treatments. Moreover, Pinel often drew on, and even used, the ethical works of the Stoics and other Ancient Greek and Roman philosophers as the basis for explaining and treating affective disorders. For Pinel, then, moral explanations were essential to explaining affective psychopathology. As Charland explains:

"…according to Pinel, affective psychopathology must employ and resort to the explanatory 'moral' vocabulary of the passions in order to correctly and accurately record these causes of alienation. This is one way in which considerations of value and morality form part of his affective psychopathology: the language in which these phenomena are couched – 'moral' terms and notions – is essential to proper explanation and description at the level of the causes of mental alienation. Simply speaking in terms of natural causes and physical states of the body is not sufficient" (Charland, 2008 p. 25).

---

[64] "Part of the difficulty here lies in ambiguities that surround the historical term 'moral'. Crichton and Pinel both agree that the passions are 'moral' phenomena in the sense that they are partly psychological. This is how the term 'moral' is usually understood in these medical discussions. However, the passions can also be considered to be 'moral' in a narrower, ethical sense. This is where the disagreement lies. Pinel argues that psychiatry must acknowledge and integrate this ethical 'moral' dimension of passions, while Crichton says precisely the opposite" (Charland, 2008 p. 24).

While the explanatory disease versus moral-problem-in-living distinction is not well developed, I want to present three reasons for why this distinction is good to have and to develop further. First, this distinction is important because it distinguishes between diseases in the narrow sense and moral-problems-in-living as being fundamentally different with respect to how they casually relate to agents. Engelhardt has called this perspective 'methodological dualism' in an attempt to interpret Thomas Szasz's argument for why mental illnesses are not real illnesses (Engelhardt, 2004; see paper 2). Methodological dualism is the view that in medicine we are primarily concerned with problems of the body whereas in morality we are primarily concerned with problems of the agent, such as certain aspects of the self that are under their voluntary control. Exactly what distinguishes direct and indirect voluntary control is not clarified by this view. But on this view medical and moral explanations are fundamentally different in kind with respect to the role that the agent plays in addressing strictly medical versus moral problems. This distinction is good to develop for both explanatory reasons and for their consequences.

While a medical patient might be required to adhere to a treatment plan, the problem he faces is not fundamentally a problem of the self. For example, failure to produce sufficient insulin in diabetes is not something that the diabetic has direct control over. A person facing a moral problem, on the other hand, is expected to gain knowledge of the self and of right and wrong, and to change their actions or their self in some ways (and often both). Engelhardt agrees with Szasz that these two methodological approaches to the self are fundamentally distinct, as do I.

We would not want to misapply these explanations to each other, since they would lead to untrue explanations and potentially ineffective or even harmful treatments for the problems we face. For example, the question of whether and in what way addictive behaviour is involuntary or voluntary has important implications for how we frame treatments for addiction. The same follows for whether addiction, or any other condition, has moral explanations, such as a potential failure of the addict to know what is good for them or certain epistemic vices like denial. Misattributing a moral explanation to a medical condition or misattributing a medical explanation to a moral-problem-in-living both have negative consequences for explaining and addressing those issues.

Another reason to develop this distinction is to prevent the potentially harmful consequences that result from how we conceive of medical versus moral approaches. In particular, by applying harmful forms of medicalisation and moralisation as such.

The concern here is not so much that we wrongly apply medical explanations to moral-problems-in-living or moral explanations to diseases-as-involuntary. The concern here is that our conception of some forms of medicalisation and moralisation are harmful even when we apply them correctly to the conditions they are intended for. One concern about some forms of medicalisation of medical conditions could be that it necessarily disempowers medical patients by treating them as passive and not involved in their own care. One concern about moralising moral-problems-in-living could be that some forms of moralisation are so overly negative that it discourages people from seeking moral guidance. The concern here is not a misapplication of categories so much that the categories that are being applied are themselves harmful. Developing the disease versus moral-problem-in-living distinction can help us rethink these categories as such. But another solution has been to demarcate and separate medicine from morality as such.

—1.2. The disease versus moral-problem-in-living distinction to constrain medicine

The disease versus moral-problem-in-living distinction has been used to distinguish between different kinds of explanations without necessarily implying a commitment to whether or not such explanations belong under the purview of medicine. Indeed some, like Benjamin Rush, have argued that both such explanations *do* and should belong under the purview of medicine. Understood in this way, the distinction is and can be purely explanatory in nature. However, the disease versus moral-problem-in-living distinction has also been used to constrain the scope of medicine by separating it from morality (Szasz, 1961; Sadler, 2005). Indeed, it is primarily this demarcating function that has been the topic of debate regarding the disease versus moral-problem-in-living distinction. Many such debates are about whether a particular condition is a medical versus a moral problem, such as Cluster B Personality disorders (Charland, 2004). But there is the shared implication that 'purely' moral problems fall out of the purview of medicine.

In this section I will present three hypotheses to explain what motivates the disease versus moral-problem-in-living distinction as a demarcation tool for the separation of medicine from morality. I present hypotheses and not claims because it is not my purpose here to defend a particular historical thesis for what gave rise to the disease versus moral-problem-in-living distinction as a demarcation tool. These hypotheses are not important as historical theses because these hypotheses attempt to capture the continuing motivations that sustain the distinction as a demarcation tool. In the next section I will lay out the costs and consequences of using the disease versus moral-problem-in-living distinction to constraint the scope of medicine.

The first explanation of what motivates the disease versus moral-problem-in-living distinction as a demarcation between medicine and morality is the view that moral explanations are not scientifically valid and therefore do not belong in medicine. On this view, moral explanations should be kept out of medicine not because they are appropriate for moral problems but inappropriate for medical problems but because moral explanations are taken to be invalid as such. This view may be due to the perspective that believing in free will in unscientific.

Strictly speaking, however, this kind of motivation is not so much a reason for maintaining a disease versus moral-problem-in-living distinction to separate medicine from morality as much as it is a reason to *reject* such a distinction. Such a motivation would lead to a rejection of the distinction because moral-problems-in-living themselves are viewed as 'a myth'. This is Thomas Szasz's reading of the nature of psychiatry, namely, that it views voluntary and moral explanations not merely as non-medical but as invalid (see paper two). Szasz therefore argued for a medicine versus morality distinction precisely to protect moral problems from such medicalisation, on the view that psychiatry would apply involuntary and non-moral explanations to everything. But this is quite a radical reading of the assumptions of the medical profession and of medical history. While concepts such as 'free will' and 'morality' may certainly be regarded with a sceptical eye by some in the medical profession, Szasz's claims that medicine was 'at war' with the very concept of moral responsibility was an overstatement (Szasz, 1987). Rather, the hypotheses that I prefer for the explanation of the medicine versus morality distinction are more mundane, though I will argue that Szasz was right that medicine has a bias against moral and therefore voluntary explanations.

The second reason for using the disease versus moral-problem-in-living distinction as a demarcation tool for medicine and morality is in an attempt to address and prevent the kinds of problems that occur when such categories are misapplied to each other. This was explained earlier. For example, there are legitimate concerns regarding how much of a role agency does in fact play in certain problems. Misattributing such a role can be either over blaming or denying one's agency. One way to deal with such problems under the purview of medicine would be to better develop the explanatory distinction between involuntary and voluntary problems by determining which processes we do and do not have control over, and subsequently which are subject to moral guidance or not. However, developing such a distinction and applying it to cases is both difficult and wrought with the risk of blaming people, especially from the perspective of medicine. Another and 'easier' way to deal with such problems is to simply demarcate medical and moral problems into different realms altogether. That is, rather than figuring out what people actually do and do not have control over and whether medical intervention or moral guidance (or some particular mix of them) is most effective to help someone, we can bypass all of that and simply stipulate that medicine will exclusively focus on problems of an involuntary and non-moral nature. This can prevent blaming people by just taking it as a given that diseases simply are those conditions that happen to us. Subsequently, there is no real need to figure out what people do and do not have control over. This second motivation therefore does not deny that voluntary and moral explanations can be valid, but rather just avoids the whole issue of making the distinction for explanatory purposes and simply just defines medicine as dealing with involuntary and non-moral problems as such. You can see this as an attractive position both from the perspective of those wanting to prevent the moralising of medical problems, but also from those seeking to prevent the medicalisation of moral-problems-in-living, such as Thomas Szasz. However, this motivation is reactive since it is motivated by an underlying fear of misapplying the distinction.

The third kind of reason is related to the second in the sense that it is a response to either an overly narrow view of medicine or of morality. From the perspective of medicine, it is the view that morality is primarily concerned with shaming and blaming in general, and therefore moral terms and approaches should be kept out of medicine. However, moral terms and approaches should not be kept out because there is a risk that they will be misapplied, but because such terms are claimed to be inherently harmful[65]. From the perspective of morality, it is the view that medicine, and by extension medicalisation, necessarily turns all patients into helpless victims. Therefore, such 'infantilising' views need to be kept out of the realm of morality not because such medicalisation can be misapplied, but because they would be inherently harmful. Szasz has elements of this view, as can be seen with his cynical views of psychiatrists. This third kind of motivations is neutral to the question of whether moral explanations are valid or not, but it is primarily a reaction to particularly negative views of either medicalisation or moralisation. This third motivation is reactive not because it is driven by fear of misapplying the disease versus morality distinction but because it is driven by a fear of an inherently harmful conception of either medicalisation or moralisation. I will turn now to the costs and consequences of demarcation.

---

[65] See Bernard William's 1986 *Ethics and Limits of Philosophy* on the view that blame is central to morality.

—1.3 The costs of the disease versus moral-problems-in-living distinction to constrain medicine

While the disease versus moral-problem-in-living distinction as a demarcation between medicine and morality may be drawn in an attempt to prevent certain problems, such as not blaming patients, using this distinction for such purposes does not come without its costs.

One cost is that the narrow conception of disease as involuntary (a conception that the demarcation process relies on) necessarily incentivises involuntary explanations of diseases. This poses a problem when a condition that we regard as a disease turns out to involve more choice than is 'normal' for prototypical diseases. Charland argues that such a scenario poses a dilemma for providing a complete explanation for a condition that will best serve the development of a treatment for it (Charland, 2008). The problem is due to the fact that such explanations come at the expense of, or at least threaten, its disease status. However, this is exactly what qualifies a condition for medical explanation and thus certain 'non-explanatory goods' such as being free from blame, stigma, and having access to medical resources (Charland, 2008)[66]. Given that all diseases involve suffering to some extent, it seems highly unlikely that, in such a dilemma, providing a more complete explanation will be chosen over a condition losing its medical status.

This kind of dilemma can again best be seen in the debate around addiction. Choice models of addiction have received pushback in part because some people think that such models threaten the disease status of addiction and the anti-stigma campaigning and public funding that this status is based on (Heather, 2017). Not only does this dilemma make it harder to provide a voluntary explanation for addiction, it also undermines *in*voluntary explanations of addiction. Take for example the fact that, in response to repeated criticisms of the disease model of addiction (in particular evidence that addictive behaviour is not involuntary but voluntary) what it means for addiction to be involuntary has become vaguer and vaguer at the cost of its explanatory power[67].

Satel and Lilienfeld have called this dilemma a 'Faustian bargain', where calling addiction a brain disease has indeed helped increase funding for research and treatment for addiction. However, that funding has all gone to models of addiction that presume that addictive behaviour is involuntary (Satel and Lillienfeld, 2014, p. 4). For example, a lot of funding has gone into medications that attempt to lower drug cravings, with mixed results. Rather, funding could also have been funnelled into research on how to change the incentives of addictive behaviour which has been shown to be effective in contingency management programmes. Funding is thus diverted not only into less effective treatments but also diverted *away* from more effective treatments. This can be seen in the United States where the National Institutes of Health supports research that assumes the brain disease model of addiction (see Satel and Lillienfeld, 2014).

---

[66] "However, the fact that the passions were so intimately bound up with questions of value and morality created a dilemma for medical scientists of the passions. If the passions are inextricably linked with questions of value and morality, then they seem to fall outside the sphere of medical science, strictly speaking. So to include this aspect of the passions in medical science threatens the scientific credentials of that science. On the other hand, to banish them risks being grossly untrue to the phenomena. Neither option is very inviting; a true dilemma" (Chardland, 2008 p. 19).

[67] "The appeal to compulsion understood as irresistible desire is key to the orthodox conception's explanation of persistent use in the face of negative consequences….Softening the meaning of compulsion [in response to criticisms that addictive desires can in fact be resisted] costs the orthodox conception its explanatory force" (Pickard, 2016 p.10).

The disease versus moral-problem-in-living distinction to separate medicine and morality does not only create a dilemma between a condition's medical status and the ability to explain certain conditions. The distinction also creates a dilemma between disease status and attributing moral agency to patients, since such agency can threaten the condition's disease status.

Some have already pointed out that those addicts that believe in the disease model of addiction are more likely to relapse, presumably because believing that addictive behaviour is involuntary makes it less likely that you believe that you can resist using drugs (Miller *et al.,* 1996). But my point here is not to argue that the disease model of addiction *necessarily* gets the role of agency wrong in addiction. Rather, my point here is the following. If it turns out that agency does play a larger role in addiction than the disease model allows for, then incorporating that role into our models now conflicts with the rhetoric that viewing addiction as an involuntary disease helps addicts accept that they have a problem. So while the disease model of addiction may help addicts accept that they have a problem, it may possibly also make it harder for addicts to find the will to recover. To be sure, a 'will to recover' is important in virtually any medical condition (see Pearce and Pickard, 2010). But the role of choice in addiction is especially wrought and hard to understand if the disease status of addiction is partly based on addictive behaviour being viewed as involuntary. Indeed, some proponents of the disease model have acknowledged that the role of choice in addiction as a disease presents a paradox. Addictive behaviour is conceptualised as involuntary while changing such behaviour is somehow still subject to voluntary change:

"Even more paradoxical is the fact that, although the alcoholic is not at fault for having the disease of alcoholism, personal responsibility is the cornerstone in the process of recovery. There is a volitional component to recovery—to seek and accept treatment for the involitional component, i.e., the loss of control over alcohol or drug use. In order to recover from the diseases of alcoholism and drug addiction, outside help is often necessary to strengthen the volition to maintain abstinence and suppress the loss of control" (Miller and Chappel, 1991 p. 203-204).

But if the 'loss of control' over using drugs is 'involitional' in addiction, and if recovery requires volition, how then can one 'suppress the loss of control' in addiction recovery? Even accepting external help implies that the loss of control in addiction can be controlled by the addict in some sense. This shows approaches that want to have it both ways. They acknowledge agency, yet insist on involuntariness in addiction, which undermines its explanatory power.

Moreover, we can see in the history of addiction that after certain non-explanatory goods have been secured through calling addiction a disease in the narrow sense, i.e., as involuntary, there are often 'course corrections' about the extent to which it is a disease in this narrow sense.

For example, the British inebriety reformers in the 19th century argued for stricter regulation of public drunkenness on the grounds that alcoholism posed a threat to the public health (Johnstone, 2001). However, after those inebriety reform laws were passed, many in the reform movement were keen to stress that addiction being a disease does not mean that addicts do not have choice, and that denying the role of choice can be detrimental to their recovery:

“By disease is popularly understood a state of things for which the diseased person is not responsible, which he cannot alter except by the use of remedies from without, whose action is obscure, and cannot be influenced by exertions of his own. But if, as is unquestionably true, inebriety can be induced by cultivation; if the desire for drink can be increased by indulgence, and self-control diminished by lack of exercise; it is manifest that reverse effects can be produced by voluntary effort; and that the desire for drink may be diminished by abstinence, and self-control, like any other faculty, can be strengthened by exercise. It is erroneous and disastrous to inculcate the doctrine that inebriety, once established, is to be accepted with fatalistic resignation, and that the inebriate is not to be encouraged to make any effort to mend his ways. It is more so, since inebriety is in many cases recovered from, in many diminished, and since the cases which recover or amend are those in which the inebriate himself desires and strives for recovery” (1908 report p. 5-6, in Johnstone, 2001 p. 49).

Similar ‘course corrections’, or clarifications, can be seen in the rebranding of addiction as brain disease in the 1990’s, which helped secure more public funding for addiction research (see Leshner, 2010). In particular, Alan Leshner wished to clarify that, while addictive behaviour is compulsive, this does not mean that addicts have no control over their addictive behaviour:

“…the recognition that addiction is a brain disease does not mean that the addict is simply a hapless victim. Addiction begins with the voluntary behavior of using drugs, and addicts must participate in and take some significant responsibility for their recovery. Thus, having this brain disease does not absolve the addict of responsibility for his or her behavior, but it does explain why an addict cannot simply stop using drugs by sheer force of will alone” (Leshner, 2001).

But not being able to stop using drugs by ‘sheer force of will alone’ need not be explained by a brain disease. Indeed, most human problems, even voluntary ones that include learned behaviours, cannot be solved through ‘sheer force of will alone’. This was already recognised by ‘moralists’ in medicine such as Rush and Pinel who were well aware that changing behaviour and the self requires knowledge, time, and effort. Yet none of such human problems are brain diseases. This seems to suggest that the brain disease model is more about securing non-explanatory goods. To the extent that there is this tension between explanation and disease status, there is some truth to Szasz’s analysis of medicine and psychiatry, namely his claim that medical thinking is antagonistic to viewing human beings as agents facing distinctly moral-problems-in-living. While Szasz is wrong that the medical profession rejects voluntary or moral explanations, they did wish to address concerns over shame and blame by removing issues of agency and morality from medicine as much as possible. But as a result of doing that, calling something a disease necessarily *does* imply a lack of agency, and therefore applying disease concepts to conditions that involve more agency *does* mean a denial of that agency. I turn now to the hard and soft the disease versus moral-problem-in-living distinction to constrain medicine.

## —2.0 The hard and soft versions of the disease versus moral-problem-in-living distinction

I have argued that the disease versus moral-problem-in-living distinction as a demarcation between medicine and morality is primarily made in order to address two things. First, certain problems that arise from misapplying medical and moral explanations, such as blame and disempowerment. Second, in order to secure certain non-explanatory goods, such as funding for research and treatment. In part two I now distinguish between a hard and a soft version of the disease versus moral-problem-in-living distinction to separate medicine and morality. There has already been a lot of criticisms of the disease versus morality distinction as a demarcation between medicine and morality in the literature (Zachar and Potter, 2010a; Charland, 2010). This criticism is directed towards a narrow conception of disease and of medicine as primarily focused on problems of an involuntary biological nature. Such a view of disease has already been rejected in favour of a broader conception of disease defined by negative consequences (Kendell, 1975).

One could argue that the disease versus moral-problem-in-living distinction has already been rejected. I will argue, however, that it is only the 'hard' version of the distinction that has been rejected, which holds that there can and should be no overlap between medical and moral kinds. But the rejection of this hard version of the disease versus moral-problem-in-living distinction does not imply that the distinction has been rejected entirely. We have not returned to Benjamin Rush's medical-moral synthesis. Instead, we have a soft version of the disease versus moral-problem-in-living distinction, where diseases can have moral components but must also have non-moral components to count as a disease. While the soft version of the distinction is a rejection of the hard version of the distinction, it is still motivated by many of the same concerns, such as stigma and blame. Later in part three, I argue that the non-moral components in the soft version of the distinction necessarily narrow the concept of disease and that it is therefore subject to the same criticisms as the hard version. I will also argue that we should reject both hard and soft versions and address the problems that motivate them differently. To clear the ground for my own position, we first have to discuss the criticism that is levied against the hard version.

### —2.1 Criticism of the hard version of the disease versus moral-problem-in-living distinction

As we will see below, the hard version of the disease versus moral-problem-in-living distinction to separate medicine and morality has rarely been directly criticised. Rather, it has been criticised and indirectly rejected by rejecting the narrow conceptions of disease, medicine, and morality that it rests on. This was due to the larger debate about the definition of disease that started in the 1960's (see paper one). In particular, if diseases are primarily defined by some kind of negative consequence, then it doesn't matter whether a condition involves involuntary or voluntary processes, or moral components. We find more direct evidence of such criticism and rejection when we look at the debates regarding the medical status of particular conditions, such as addiction and the Cluster B Personality Disorders (CBPDs). But in the next section we will see that there still remains a soft version of the disease versus moral-problem-in-living distinction as a to separate medicine and morality as a means to prevent the misapplication of these domains.

Criticisms of medical expansion are now often framed in terms of medicine expanding beyond its scope of practice from a central core that most people agree on. But the criticisms of medicine in the 1960's were more radical in nature. Rather, the expansion of medicine into controversial areas was taken to reveal that medicine had no more claim to calling an uncontroversial condition a disease than it does to call a controversial condition a disease. In other words, all medical authority was suspect. Such criticisms could especially be seen from Michel Foucault (Foucault, 1973), and from Thomas Szasz who argued that psychiatry was not a legitimate branch of medicine at all (Szasz, 1961). Szasz argued that mental illnesses are not real illnesses and that therefore the entire field of psychiatry was an institutional farce. Szasz's criticism of the concept of mental illness was rejected primarily because his very narrow conception of illnesses as lesions was no longer in use (see Kendell, 1975). But Szasz's criticism of psychiatry did raise deeper questions of how diseases and medicine should be defined.

There were four issues in particular that needed to be addressed to give modern medicine and psychiatry legitimacy in response to Szasz. The first was the issue of how diseases were to be defined if it was not going to be defined by a specific biological aetiology, such as lesions. This question was of importance to psychiatry because it was still dominated by psychoanalysis until the 1960's. Due to psychoanalysis' dominance, psychiatry was already coming into conflict with other models for mental illness such as behaviourism and biomedical models of mental illness (Spitzer & Endicott, 1978). There was therefore a need to give a definition of illness and mental illness that did not necessarily take a side between these different schools of thought. The second issue had to do with the roles of facts and values in distinguishing disease from non-disease. In particular, a major concern in defining disease was, on the one hand, to address criticisms that the concept of disease was nothing more than a reflection of medicine's or society's values. On the other hand, to reflect the fact that a patient's individual values are relevant in some sense to defining what their health priorities are in a medical setting (Boorse, 1975; Wakefield, 1992).

The third and fourth issues concern specific kinds of cases where these normative issues come up. Namely, how does medicine distinguish between medical and normal suffering, and how can one distinguish medical pathology from mere social difference and deviation.

Elsewhere I have argued that a particular issue that concerned Thomas Szasz regarding the concept of illness and mental illness was the medicine versus morality distinction (see paper two). In particular, Szasz was concerned that the very notion that medicine should be primarily concerned with problems of an involuntary nature was being rejected by the concept of mental illness. After all, Szasz thought that the concept of mental illness implied a rejection of the medicine versus morality distinction. However, his concern was not so much that medicine would start to include voluntary and moral explanations. Rather, his main concern was that medicine would expand to include problems of a voluntary and moral nature without explaining them in voluntary and moral terms. In particular, he was concerned that medicine would attempt to explain and treat problems of such a nature through involuntary means, and involve involuntary explanations of the condition in question and therefore coercive treatments. I have argued that this concern of Szasz was more or less neglected in relation to the other four issues that I raised above. I will return to the implications for the medicine versus morality distinction shortly.

While there remain many disagreements about how to define the concept of disease and illness to this day, a certain kind of consensus has emerged in response to Szasz's criticisms of psychiatry. First, the lesion conception of disease was rejected, and some even argued that this conception of disease was no longer in use by the time of Szasz's criticism of psychiatry (Kendell, 1975; Pies, 1979). It was also argued that medical concepts such as disease or illness should not be defined by a specific underlying cause, i.e., a particular aetiology. Rather, diseases should be defined by their negative consequences to the patient (Scadding, 1967). Initially, such negative consequences were defined in terms of 'biological disadvantage', such as reduced reproductive success and survival rates (Kendlell, 1975). However, such negative consequences did not capture the fact that many illnesses didn't reduce one's reproductive success or survival rates, but rather caused intense suffering and impairment in one's daily living. In particular, the concept of illness and harm became important notions to capture the fact that, from a medical perspective, what we are concerned about is whether or not a condition is harmful to the patient (Boorse, 1975; Wakefield, 1992). The concepts of illness and harm were also helpful in addressing issues of distinguishing between pathology and mere difference. If a condition was a deviation from population norms but it did not cause harm then it was merely a difference and not a disease (Wakefield, 1992). There are still major disagreements about how to distinguish medical from non-medical 'normal' forms of harm in the concept of disease debate to this day (see Schuur, 2019). For example, there are disagreements about the relation between value-free facts and value-laden norms in distinguishing between disease and non-disease. Nevertheless, there is a consensus view that diseases should not be defined by a narrow aetiology but should instead be defined by some kind of negative consequence that can somehow be distinguished from non-medical negative consequences. One of the proposals for making such a distinction is by introducing a second component to the concept of disease, such as dysfunction (Boorse, 1975; Wakefield, 1992). Diseases, or medical disorders, are then dysfunctions that harm, i.e., harmful dysfunctions.

The implications of this rejection of the narrow conception of disease on the view that medicine should not be concerned with moral-problems-in-living is rarely made explicit. Because the conception of disease as involuntary is an instance of a particular aetiological view of disease, the rejection of this narrow conception of disease in favour of a broader conception of disease should constitute a rejection of the view that medicine should be constrained to involuntary non-moral conditions. On this definition, something is a disease regardless of the role of agency or morals in it. The fact that the broad conception of disease is an implicit rejection of the disease versus moral-problem-in-living distinction to separate medicine and morality is most often acknowledged within the context of debates regarding the medical status of particular conditions. Marc Lewis, for example, distinguishes between the narrow conception of disease as involuntary and the broad conception of disease as disability (Lewis, 1991). Lewis points out that the narrow conception of disease as involuntary would exclude conditions of a voluntary and moral nature:

"If the etiology and/or the course of the condition appears to be primarily under the control of the individual, then it is thought to be a moral problem. If, on the other hand, the etiology and/or course the condition is primarily beyond the control or intention of the individual, it is more acceptable to call that condition a disease" (Lewis, 1991 p. 256-258).

Lewis uses this distinction to argue that addiction is a disease in the narrow sense, i.e., an involuntary and therefore non-moral condition (Lewis, 1991 p. 258). However, Lewis then goes on to argue that whether or not addiction is a disease in the narrow sense of the term has become irrelevant because of the acceptance of broader conception of disease. If diseases refer to disabilities then addiction is uncontroversially a disease regardless of whether it is voluntary or involuntary in its aetiology or explanation (Lewis, 1991 p. 258). While not made explicit by Lewis, the broader conception of disease as disability clearly allows for the existence of diseases of a voluntary and moral nature because they are irrelevant to whether something is a disability.

This point is made explicitly by Wakefield: "…a brain disorder may in part simultaneously constitute a moral defect" (Wakefield, 2017 p. 65). In response to another Marc Lewis arguing *against* the brain disease model of addiction, Wakefield argues that whether or not addictive behaviour is involuntary or not, or a 'moral defect' or not, is irrelevant to the question of whether addiction is a medical disorder (Lewis, 2015; Wakefield, 2017 p. 55). Whether a condition is involuntary or voluntary, non-moral or moral, are issues of causation and aetiology that have been made irrelevant by defining diseases by their negative consequences. If a condition is a harmful dysfunction it is a medical disorder, even if it involves a 'moral defect'. For Wakefield, then, addiction is a medical disorder regardless of whether it is a 'moral defect' or not. The voluntary and moral components of addiction should have no bearing whatsoever on its being a medical disorder for Wakefield just as long as addiction is a harmful dysfunction (Wakefield, 2017 p. 65).

The rejection of using the disease versus moral-problem-in-living distinction to separate medicine and morality is also made explicit in the debate over the medical status of Cluster B Personality Disorders (CBPDs). Charland has argued that CBPDs should be regarded as 'moral' rather than 'clinical' conditions (Charland, 2004). The CBPDs refer to histrionic, borderline, and anti-social personality disorders in the DSM-5 (APA, 2013), some of which are defined in explicitly moral terms. Anti-social personality disorder, for example, is described as involving a 'lack of empathy' and disrespect 'for the rights of others', while borderline personality disorder is described as involving 'manipulation' of others. Charland moreover argues that many of the treatments for the CBPDs involves a 'moral conversion' to be a better person. Charland points to aspects of Dialectical Behavioural Therapy (DBT) as involving such moral conversion, such as the therapeutic alliance, and agreements to swear off manipulative behaviour and to be more honest. Charland argues that both the indispensable moral features of the CBPDs and the moral components of their treatment means that we should regard CBPDs as 'fundamentally moral' rather than 'clinical' conditions (Charland, 2010 p. 122). One implication of this is that physicians should not have an exclusive claim on the treatment of CBPDs because non-physicians like religious counsellors might have expertise in providing moral guidance (Charland, 2004 p. 74).

Charland's arguments have raised debates not only with respect to the nature of CBPDs, but also for whether or not and to what extent conditions of a moral nature should fall under the purview of medicine. Some critics have argued that, if CBPDs involve moral components, then this does not necessarily imply that CBPDs are thereby moral and not clinical conditions (Zachar and Potter, 2010a). Indeed, if being a medical disorder involves some kind of dysfunction and/or harm, then that is compatible with the disorder also involving significant moral components.

Zachar and Potter have also rejected not only the narrow conception of disease as referring to involuntary and non-moral conditions but have also rejected the narrow conception of moral problems as primarily referring to actions. Instead, they view moral problems as also referring to other aspects of the self, such as one's character (of which we would have some voluntary control over) such as emotional functioning and cognition (Zachar and Potter, 2010a).

Charland has responded to these kinds of criticisms by clarifying that he is not arguing that CBPDs are not clinical conditions, but that CBPDs should be regarded as fundamentally and primarily moral conditions rather than clinical conditions (Charland, 2010)[68]. Charland therefore claims not to hold a hard medicine versus morality distinction, but only claims that some conditions, like CBPDs, should be regarded more as moral conditions than as medical conditions. What Charland seems to be doing is conflating and confusing the disease versus moral-problem-in-living distinction as an explanatory distinction versus as a demarcation between medicine and morality as such. Charland is arguing that CBPDs are moral in nature and therefore require moral explanations but that CBPDs are still diseases in the broad sense of the term and therefore medical professional should still be involved (though not exclusively). This is why it is important to be clear whether we are applying the disease versus moral-problem-in-living distinction as an explanatory distinction or as a means to separate the domains medicine and morality. Charland seems to be primarily concerned with opening up medicine to non-physicians and explanations that are not narrowly medical. He is not arguing that medicine should focus exclusively on non-moral problems nor that non-physicians should exclusively treat problems of a moral nature. Charland, who first appears as a defender of the view that the domains of medicine and morality should be kept apart, also rejects this view and instead allows overlap between the domains[69].

The hard disease versus morality distinction is therefore widely rejected when it comes up regarding particular conditions, such as CBPDs. Moreover, the reason why the hard version of the disease versus morality distinction is rejected is because it is based on a narrow conception of diseases as involuntary and moral problems regarded as 'wrong actions'. These views themselves are rejected in favour of broader conceptions both of disease and of moral problems of living.

From one perspective, this criticism of the hard medicine versus morality distinction could be regarded as a complete rejection of using the disease versus moral-problem-in-living distinction to separate medicine and morality. However, I will argue that this does not mean that the disease versus moral-problem-in-living distinction is entirely rejected to separate medicine and morality. Indeed, I will go on to argue that there is still a soft version of this disease versus moral-problem-in-living distinction as means to separate medicine and morality, and that it is motivated the same concerns as the hard version of that distinction. I will clarify the nature of this soft version of the disease versus moral-problem-in-living distinction in the next section before turning to providing arguments for why we should reject this version too in part three.

---

[68] "I argue that borderline personality disorder and the other DSM-IV cluster B personality disorders are not really medical, or clinical, kinds at all. They are 'really'—that is, *fundamentally*, but *not exclusively*—moral conditions for which, ultimately, there is apparently no medical cure" (Charland, 2010 p. 122 original emphases).

[69] "I do not adhere to the view that 'medical' or moral kinds (or alternately, 'clinical' or moral kinds) are mutually exclusive" (Charland, 2010 p. 120).

—2.2 The soft disease versus moral-problem-in-living distinction to constrain medicine

In response to Charland's claims that CBPDs are moral and not medical kinds, Zachar and Potter draw on John Sadler's work on medical and moral kinds to argue for a soft version of the disease versus moral-problem-in-living distinction (Zachar and Potter, 2010a). Here I will present Sadler's soft disease versus moral-problem-in-living distinction and his arguments for it. In part three I will present arguments for why we should also reject this soft version of the distinction.

One of Sadler's scholarly concerns is the distinction between medical disorders and problems of a moral nature, in particular, criminal, immoral, and unconventional behaviour, the so-called 'mad versus bad' distinction (Sadler, 2005 p. 210). It should be noted that Sadler's conception of moral problems of living is much narrower than what I have described so far, which are moral-problems-in-living of figuring out what actions, habits, and values we require to develop ourselves and flourish. The kinds of phenomena Sadler is concerned with are not moral problems in this sense, but rather refer to only a small fraction of what in my view the whole set of of moral-problems-of-living consists of. The mad versus bad distinction is important for Sadler for two reasons. First, because it is sometimes difficult to tell whether a problem is a medical disorder or a 'vice', as he calls it. Second, because the distinction is important to defining and demarcating the nature and boundaries of medicine as opposed to morality and the law (Sadler, 2008 p. 1).

In seeking a clearer distinction between medical disorders and vices, Sadler wants to be able to avoid two kinds of problems in particular (Sadler, 2005 p. 220). The first problem is the phenomenon of regarding a condition as a medical disease only because it is regarded as immoral by others, whether implicit or explicit, such as homosexuality used to be. The problem that Sadler wants to avoid does not necessarily concern cases where we wrongly medicalise a condition that should in fact be moralised, i.e., confusing a moral problem with a medical problem. Rather, the problem that Sadler wants to avoid is wrongly medicalising a condition that is *also* wrongly moralised, as most contemporary scholars now view to be the case with homosexuality. Rather, homosexuality is neither regarded as a medical condition nor a moral problem but merely a variation in human sexual orientation. What Sadler wants to avoid, not unlike Thomas Szasz and other past critics of psychiatry, is using medical concepts as a means to medicalise problems that society regarded as immoral to begin with (Szasz, 1961). But unlike Szasz and earlier critics of psychiatry, Sadler does not think that the solution is 'abolishing psychiatry' or regarding psychiatry as not being legitimate. Sadler thinks this problem can be avoided by having a clearer disease versus vice distinction to separate them (Sadler, 2005 p. 220).

The second kind of problem that Sadler argues that a disease versus vice distinction should solve is avoiding the denial that a condition is a legitimate disease because it is wrongly regarded as immoral. For example, certain forms of insanity may in the past have been regarded as immoral or criminal when those people are actually suffering from a disease. Regarding these conditions as immoral and therefore not medical denied these people the benefits of medical treatment and excuses. Medicalisation can therefore have both benefits and harms, and Sadler wants to clarify and develop the disease versus vice distinction so that the benefits and harms of medicalisation are correctly distributed and avoided, respectively, for genuine medical problems.

As Sadler poses the question: “How can we prevent both kinds of evaluative mistake, that of false-positive and false-negative diagnostic characterisation?” (Sadler, 2005 p. 222).

It is interesting to note, however, that Sadler’s primary concern is focused on either avoiding the harms of unnecessary medicalisation or avoiding the loss of the benefits of medicalisation. Both of the problems that Sadler wants a disease versus vice distinction to help prevent concern cases where a condition is wrongly moralised and therefore either receives the harms of wrongly being medicalised, such as homosexuality, or receive the harms of not being medicalised, as in the case of people with severe mental illnesses that are wrongly moralised.

Sadler does not consider the problem of wrongly medicalising a condition because it should in fact be moralised but isn’t moralised. This was Szasz’s primary concern regarding the disease versus moral-problem-in-living distinction (see paper two). The fact that Sadler is not concerned with this kind of problem reflects in part his narrow conception of moral-problems-in-living as mostly referring to immoral behaviour, and relatedly to a widespread belief that moralising a condition is inherently harmful. That is, the belief that moralisation does not simply refer subjecting a problem to moral evaluation and guidance, but as referring to more specific and harmful practices, such as blaming and shaming people. I will return to this issue in part three.

While Szasz and Sadler were concerned with different problems, they both want a disease versus moral-problem-in-living distinction to prevent the medicalisation of moral-problems-in-living and the moralisation of medical problems. But unlike Szasz, Sadler acknowledges that in many cases a hard distinction between diseases and moral-problems-in-living is impossible. Sadler argued that conditions with moral components should still count as medical disorders. He had two requirements that a moral condition must have in order to qualify as a medical disorder:

“(1) The category or construct must have testable features (clinical, etiological, predictive) that are *non morally bad* —not neutral, not morally bad, but nonmorally bad, as with any ‘typical’ disease. Etiological features in this case must be causally connected to nonmorally bad clinical or predictive features. Such linkage would ordinarily be present in etiological considerations of valid categories.

(2) Such nonmorally bad features of the category must be validly associated with the construct; for example, the nonmorally bad features of the category must integrate or cohere into a construct validation process” (Sadler, 2009 p. 223).

Important to note here is that Sadler is re-introducing an aetiological consideration for whether or not a condition is a medical disorder. That is, while Sadler does not define valid medical disorders in exclusively aetiological terms, as Szasz did, Sadler does consider certain aetiologically factors, i.e., causes, to be relevant to distinguishing disease from moral problems.

Sadler does not say much about what distinguishes medical from moral aetiologies, but one view on this question that came out of the debate about the status of CBPDs was that the role which agency plays in a condition might be relevant in determining whether a particular condition, or parts of it, can be regarded as moral or non-moral (see Zachar and Potter, 2010a).

Based on these two criteria, Sadler argues that there are three categories of legitimate medical disorder categories in their potential relation to the moral features that they might have:

"These criteria would permit three 'evaluative' kinds of (legitimate) mental disorder: completely nonmorally bad conditions, mostly nonmorally bad conditions, and mostly morally bad conditions with indispensable nonmorally bad features" (Sadler, 2005 p. 223)

Here we see that, for Sadler, a medical condition may have necessary moral components just so long as it also meets the general criteria for being a disease. In particular, Sadler argues that such conditions must have 'indispensable non-morally bad features' other than distress and suffering (Sadler, 2005 p. 223). Sadler, however, is not clear on what these features should be that will actually help us distinguish between diseases and vices. There are two options in the literature for making such a distinction, either a dysfunction criterion or an aetiological criterion. In part three I will argue that these criteria are not compatible with a broad conception of disease. As mentioned earlier, Sadler wants to have these kinds of criteria in order to prevent conditions which are not diseases from being wrongly medicalised and prevent conditions which are diseases from being denied medical status due to being moralisation. Again note that there is no concern for Sadler regarding the medicalisation of something which should be moralised. Sadler also gives five more reasons for why we need to maintain a soft disease versus moral-problem-in-living distinction (other than avoiding the two misapplication problems above) (Sadler, 2005 p. 224-226):

-1. Medicine should be concerned with health, not 'policing immoral behaviour'.
-2. Morally bad aspects in diagnoses are contentious, thus motivating anti-psychiatry arguments.
-3. Morally bad aspects are not essential for disorder status.
-4. Medicine should not directly meddle in issues of moral and/or legal responsibility.
-5. Reduction of stigma of mental disorders, disassociate them from immoralities.

It is worth noting that many of Sadler's motivations for maintaining a soft disease versus moral-problem-in-living distinction are the same motivations behind the hard disease versus moral-problem-in-living distinction to separate medicine and morality. The difference between them is the extent to which diseases and moral problems can be separated. But what they fundamentally agree on is that medicine should be focused on diseases and not moral problems. Sadler does acknowledge that such a separation may have negative consequences for those conditions not sufficiently regarded as medical enough[70]. I will return to this issue that Sadler raises after criticising the soft disease versus moral-problem-in-living distinction in part three.

---

[70] "However, it should be acknowledged that such a change in assigning disorder status for conditions will occasion other 'value consequences' which will be controversial and perhaps undesirable. Presumably, disqualifying morally bad conditions as disorders will, diagnostically, move them to the V-code status as conditions that may be a focus of treatment but are not disorders, or, possibly, move them out of the manual altogether. Those 'patients' who have historically had treatment offered to them may experience constraints (financial, practical, organisational) in accessing mental health care. Clinicians who treat the disqualified conditions may suffer from the effects of reduced reimbursement" (Sadler, 2005 p. 226).

## —3.0 Rejecting the soft disease versus moral-problem-in-living distinction as demarcation

In part three I will present my arguments in favour of rejecting the soft disease versus moral-problem-in-living distinction to constrain medicine and offer some alternative responses to the issues that motivated the distinction. My main argument for rejecting the soft version of the distinction is that it is still subject to the same criticisms of the hard version. I argue that the soft disease versus moral-problem-in-living distinction distinction to constrain medicine is incompatible with the broad conception of disease and that it incentivises involuntary explanations. Furthermore, I argue that the soft distinction imposes an unhelpful framework for addresses legitimate issues like stigma. But first I will briefly summarise why we are right to reject the hard version of the disease versus moral-problem-in-living distinction as a means to constrain medicine and why proponents of the soft version of the distinction think that their version of the distinction is not subject to the same criticisms as the hard version.

The hard disease versus moral-problem-in-living distinction to constrain medicine depends on a narrow conception of disease as necessarily involuntary and non-moral, which are aetiological criteria for being a disease (Lewis, 1991). Moreover, this narrow view of diseases as defined by a particular aetiology has been rejected within the wider debate about the meaning of disease (Boorse, 1975; Wakefield, 1992). This is because such a narrow conception fails to capture the fact that what is most important about diseases is not their causes, but their negative consequences (Kendler, 1975). The direct rejection of the narrow conception of disease is therefore an indirect rejection of the hard disease versus moral-problem-in-living distinction to constrain medicine. A broad conception of disease which allows diseases to have voluntary or moral components is therefore also a rejection of the hard disease versus moral-problem-in-living distinction to constrain medicine. The soft version of the medicine versus moral-problem-in-living distinction is soft because it presents itself as a rejection of the hard version of the distinction (and the narrow conception of disease that the hard distinction rests on) while still seeking to maintain the strategy of separating medicine and morality. This is visible in the way in which Zachar and Potter criticise Charland's claims that CBPDs are moral and not medical conditions by arguing that this is a false distinction. Their argument holds that we cannot make a hard distinction between narrowly medical and moral problems once we accept a broad conception of disease as defined by negative consequences (Zachar and Potter, 2010). Indeed, some conditions that have medically significant harm may well turn out to also have irreducible moral components.

Yet, the issues that motivated the hard distinction to constrain medicine are not automatically addressed by this broader conception of disease, such as how to address issues of blame, shame, stigma, and how to prevent the misapplication of medical and moral categories. The soft version of the disease versus moral-problem-in-living distinction still tries to remove purely moral-problems-in-living from medicine. It does this by stipulating that any condition that turns out to be a moral-problem-in-living must also have an 'indispensable non-moral component' that also makes it a disease in the broad sense of the term. In other words, the soft version of the distinction is premised on the assumption that we can retain this broad conception of disease while still keeping out purely moral-problems-in-living from the domain of medicine.

## —3.1 Why we should reject the soft disease versus moral-problem-in-living distinction

The soft disease versus moral-problem-in-living distinction assumes that we can distinguish between diseases and moral problems while not reverting to a narrow conception of disease. The soft distinction aims to retain our broad conception of diseases as defined primarily by their negative consequences and not by their causes. In this section I will argue that the soft distinction *necessarily* leads to a narrowing of our conception of diseases by either regarding diseases as necessarily involuntary or regarding involuntary conditions as more 'disease-worthy'. Therefore, the soft version of the distinction values involuntary explanations over others.

Defining diseases solely in terms of their negative consequences is regarded as too broad a conception of disease because it fails to distinguish the negative consequences associated with diseases from non-disease, such as normal suffering (Kendler, 1975; Boorse, 1977). One solution to this kind of problem is to add a limiting criterion to disease, such as an underlying dysfunction, which distinguishes those harmful consequences associated with diseases from those associated with non-diseases (Boorse, 1975; Wakefield, 1992). One potential way to establish a soft disease versus moral-problem-in-living distinction is to argue that conditions with moral components must have some non-moral features that make them diseases other than simply distress and suffering (Sadler, 2005). Dysfunction has been offered as providing such a non-moral criterion to distinguish diseases with moral components from non-disease moral problems (Zachar and Potter, 2013).

Dysfunction is not an aetiological concept but a broadly-construed normative concept (Wakefield, 1992). A dysfunction tells us that there is a proper function regarding a biological system or sub-system and that this function is somehow impeded or failing. Dysfunction as a concept is neutral with respect to the causes of the dysfunction: it does not tell us why a certain function is not being performed, only that it is not being performed (Boorse, 1977). Dysfunction appears to provide a non-moral feature by which to determine whether a condition that has moral features is also a disease or whether it is not a disease at all (Zachar and Potter, 2010a). Moreover, the dysfunction criterion does not impose a particular aetiological criterion on some condition being a disease. We see the use of this criterion to establish a soft disease versus moral-problem-in-living distinction in the work of Zachar, who argues that CBPDs are medical disorders because they involve dysfunctions and are therefore medical disorders (Zachar, 2011).

Using dysfunction as a means to establish a soft disease versus moral-problem-in-living distinction assumes that dysfunctions are necessarily non-moral in nature (Sadler, 2005 p. 223). But this assumption is flawed. Because the concept of dysfunction is neutral with respect to aetiology, it does therefore not exclude particular kinds of aetiologies, including voluntary or moral aetiologies. It may therefore turn out that some failures of function, i.e., dysfunctions, can in principle be voluntary and moral in nature. The usage of the dysfunction concept as a non-moral criterion that diseases must have therefore appears to assume that dysfunctions are inherently non-moral in nature, or it accepts that dysfunctions can be moral in nature but that it is only the non-moral dysfunctions that are the kinds of dysfunctions that diseases are required to have. The former is question-begging, and the latter raises the question of why we should deny or discount dysfunctions as being medical simply because they are voluntary and moral in their aetiologies.

Because dysfunctions can be moral in nature, dysfunction alone is insufficient to serve as a criterion to establish a soft distinction between diseases and moral-problems-in-living. Another criterion is therefore required to distinguish between the dysfunctions that are non-moral in nature and those that are moral in nature. In the debate regarding the moral nature of the cluster B personality disorders (CBPDs) it has been suggested that the causal role of agency within the aetiology of CBPDs can help to determine whether they qualify as a medical disorder[71]. This criterion is similar to Marc Lewis' distinction between diseases and moral-problems-in-living[72].

Indeed, conditions that we have little or no control over are often regarded as less moral in nature because moral evaluation and guidance only apply to those things we have some control over. Based on this criterion for what counts as a moral-problem-in-living, a condition should be regarded as 'not moral' or 'less moral' if the aetiology of the condition involves no or less agency than a prototypical moral problem. However such distinction is made, introducing such a criterion for diseases should raise serious concerns whether such a criterion is compatible with our broad conception of diseases as defined by their negative consequences, instead of their causes.

Answering this latter question in part depends on what is meant by involuntary and less or no control. There is one way that involuntariness could be defined that is compatible with our broad conception of disease as defined primarily by their negative consequences. A condition could be involuntary in the sense that any enduring state is not subject to direct voluntary change. On this conception of involuntary, only actions are voluntary. All enduring states on this view are involuntary insofar as part of what it means for a condition to be a condition is for it to be an enduring state, and not a discreet and isolated action that you could have chosen to bring about or not. However, this notion of involuntariness is too broad to serve as a basis to distinguish between moral and non-moral dysfunctions, because many moral problems also involve enduring states, such as vices, and aspects of the self that can only be changed over time, like character.

A more narrow definition of involuntariness would be behaviours that are not open to voluntary change over time, or a condition that involves significant involuntary processes in its aetiology even if some of its behaviours are partly voluntary. Indeed, this is how Zachar and Potter use the term as a means to determine whether a particular condition falling under the CBPD category is a medical disorder or not (Zachar and Potter, 2010b). Some medical disorders do involve involuntary behaviours, and part of what makes them disorders is precisely that their involuntary nature makes them harmful and indicative of an underlying dysfunction. The involuntary shaking in Parkinson's, for example, is part of what makes the condition clinically a harm *and* is indicative of some underlying disorder. Zachar and Potter may therefore be right that, if the behaviours described in the Cluster B Personality Disorders are involuntary in a more narrow sense, then CBPDs may involve a non-moral dysfunction and therefore be classified as proper diseases. Such a feature may therefore help to distinguish moral from non-moral dysfunctions.

---

[71] "Lack of control, or compulsivity, may itself help to legitimize disorder status" (Zachar and Potter, 2010a p. 108).

[72] "If the etiology and/or the course of the disease appears to be primarily under the control of the individual, then it is thought to be a moral problem. If, on the other hand, the etiology and/or course the condition is primarily beyond the control or intention of the individual, it is more acceptable to call that condition a disease" (Lewis, 1991 p. 256-258).

However, while such a criterion could distinguish between moral and non-moral dysfunctions, it necessarily reintroduces an aetiological criterion for medical disorders. That is, if moral versus non-moral kinds distinguish between different causal and normative explanations of the role of agency and other aspects of the self that are subject to voluntary control over time, then such a distinction necessarily undermines one kind of causal explanation in favour of involuntary and non-moral explanations. Such a criterion would clearly be incompatible with our broad conception of disease as primarily defined by their negative consequences and not defined by some particular aetiological criterion. To the extent that the broad view of disease was meant to create parity between different aetiological views of diseases, introducing an aetiological requirement for being a disease would be in tension with the broad conception of disease if not an outright rejection of this principle. To the extent that the soft disease versus moral-problem-in-living distinction to constrain medicine reintroduces an aetiological criterion for disease, this distinction is at odds with the principle of aetiological parity. Such a distinction undermines the principle of aetiological parity in medicine by carving out an exception to it. Such an exception would mean that: even if a condition involves a harmful dysfunction, it would be rejected or downgraded as a disease because it is not sufficiently involuntary and non-moral in nature.

We should reject distinguishing between moral and non-moral dysfunction on the basis of some underlying aetiological difference between them for the same reason that we should reject any aetiological criterion for a condition qualifying as a medical disorder. The primary reason for rejecting any aetiological criterion for a condition qualifying as a medical disorder is that what is most important about a condition qualifying as a disease is that it involves medically significant harm and impairment. It shouldn't make a difference if the underlying causes or nature of the condition in question involves voluntary or moral components or not. One reason why a condition involving voluntary or moral components may appear to be inversely related to a condition's severity is the assumption that if it involves voluntary or moral components that the condition is therefore easier to treat than an involuntary non-moral condition. But this assumption is wrong. Some conditions that involve learned behaviours, such as personality disorders, may be much harder to treat than certain involuntary conditions. There is therefore no necessary relationship between the clinical severity of condition's harm, the underlying dysfunction, to the role that voluntary and moral components may play in the aetiology of the condition in question.

The soft disease versus morality distinction therefore appears to necessarily reintroduce an aetiological criterion for being a disease to the extent that the only way to distinguish between moral and non-moral dysfunctions is the determination of the role that agency plays in a condition. Even if a condition is primarily regarded as a medical disorder because it involves a harmful dysfunction, the inclusion of an aetiological criterion necessarily prioritises some medical disorders as more worthy of their medical status than others. The soft disease versus moral-problem-in-living distinction to constrain medicine should therefore be rejected because it is incompatible with our broad conception of disease as defined primarily by their negative consequences. In the next section I will show that maintaining a soft disease versus moral-problem-in-living distinction is not only incompatible with our broad conception of disease but that it necessarily creates a harmful trade-off between better explanations and medical status.

3.2 The soft disease versus moral-problem-in-living distinction, a harmful trade-off

Reintroducing an aetiological criterion for disease, such as a non-moral aetiology, means that some harmful dysfunctions could either be rejected or discounted as genuine medical disorders purely because of their specific aetiologies. It would mean that two conditions that are the same with respect to their underlying dysfunction and harmful consequences may in fact not equally count as diseases if it turned out that one of those conditions involve voluntary and moral components that make it a moral dysfunction (compared to a condition that was involuntary and non-moral). This is not much different than mandating that diseases must be involuntary to be diseases, and is therefore similar to the hard disease versus moral-problem-in-living distinction.

In practice, however, reintroducing an aetiological criterion for a condition being a disease is unlikely to result in actually denying any condition their medical status. If it turned out that a particular condition currently under the purview of medicine, such as addiction, was actually a moral dysfunction, it is highly unlikely that medical professionals would argue for its expulsion from the category of medical disorders. This is because the high levels of suffering and impairment qualifies addiction as a disease regardless of the nature of the dysfunction involved.

What is more likely to happen is one of two things. First, any discovery of voluntary and potential moral explanations for certain conditions would be either denied or translated into non-moral terms. Such denial or translation would occur because voluntary and moral explanations would pose a threat to the medical status of a condition, since we require a soft disease versus moral-problem-in-living distinction to constrain medicine. This can be seen in Peter Zachar's attempts to justify the medical status of CBPDs by translating the moral terms into non-moral terms of dysfunction and harm (Zachar, 2011). Second, involuntary and non-moral explanations of a condition might be exaggerated to make sure that its medical status is not in question. This is most likely to occur with conditions whose medical status itself has been put in question.

One example where this might happening is in the debates about addiction. Addiction's status as a medical disorder was and still in large part is justified on the claim that the disease model of addiction shows that addiction is not a moral-problem-in-living (see Leshner, 1997). This argument, that addiction is not a moral problem and therefore a disease, has historically been based on the view that addictive behaviour is in some sense involuntary (see paper three). Choice models of addiction that challenge the disease model of addiction's view that addictive behaviour is involuntary therefore raise concerns that choice models imply a 're-moralisation' of addiction. Proponents of choice models, however, have argued that their models are not moral models (Heyman, 2009). Despite this, choice models still have an extra burden for being accepted. This burden is due to two interrelated reasons. First, because disease models of addiction present themselves as the only alternative to moral models of addiction, any criticism of the disease model of addiction by proponents of the choice model of addiction is seen as a pathway to re-moralise and de-medicalise addiction. Second, proponents of choice models of addiction have asserted that their models are not moral models without actually explaining why this is the case. Indeed, Heyman's own choice model of addiction, while claiming to be non-moral, draws on insights that appear to be moral, such as the importance of 'private rules' (Heyman, 2009 p. 166).

Clearly, the discovery that addiction involves a moral dysfunction would be unwelcome news, despite this changing *nothing* about the severity of the dysfunction and harm involved. Indeed, the notion that addiction may have moral components is historically regarded as a threat to its medical status (Miller and Chappell, 1991). So regardless of how the debate about the nature of addiction turns out, there are clearly incentives to insist on addiction's non-moral status. The disease versus moral-problem-in-living distinction, whether this distinction is soft or hard, therefore appears to creates incentives against voluntary and moral explanations in medicine as such. It moreover leads to regarding certain aetiologies as more 'disease worthy' than others.

Because the soft disease versus moral-problem-in-living distinction conceptually prioritises involuntary and non-moral explanations and creates incentives to undermine the role of choice in medicine, the soft disease versus moral-problem-in-living distinction should be rejected as a means to constrain the scope of medicine. In other words, the soft disease versus moral-problem-in-living distinction does not succeed in addressing the problems facing the hard version of the disease versus moral-problem-in-living distinction. The hard version was rightly rejected in the first place, namely, for prioritising a narrow aetiology for being a diseases over the negative consequence of diseases and for disregarding the role other causes can play in certain conditions. The soft version of the disease versus moral-problem-in-living distinction simply replicates these two problems while attempting to argue that problems of a moral nature can in fact also be diseases. But if a requirement for being a disease is having 'indispensable non-moral features', then such non-moral features need to be clarified. What we have seen, however, is that such non-moral features are incompatible with a broad conception of disease where aetiologies are regarded as equal with respect to disease and medical status. Moreover, implementing such non-moral features as criteria of diseases would necessarily lead to either undermining or denying the role of agency and moral features in some conditions, and could lead to over-emphasising or exaggerating the involuntary nature of certain conditions in order to secure their medical status.

Before turning to my last argument against the soft disease versus moral-problem-in-living distinction, it is worth noting how my approach to the medicine's relation to morality differs from Thomas Szasz (Szasz, 1961). In an earlier paper (see paper two) I argued that Szasz regarded diseases are primarily involuntary and non-moral in nature and by extension regarded such conditions as being the primary problem that medicine should treat. Szasz's primary concern, I argued, was that medicine was treating more and more problems that are voluntary and moral in nature as involuntary and non-moral conditions. Szasz's solution was to re-affirm a clear disease versus moral-problem-in-living distinction to prevent the medicalisation of moral-problems-in-living. Szasz takes for granted that diseases are necessarily involuntary and non-moral in nature and I argued that diseases are instead defined broadly by their negative consequences. But to the extent that there are incentives against introducing voluntary and moral explanations and incentives in favour of involuntary and non-moral explanations in medicine, Szasz is right that medicalisation often can and does imply a denial of agency. But unlike Szasz, my solution is not to constrain the scope of medicine to non-moral problems. Instead, my solution rejects such an approach in favour of fully allowing for voluntary and moral explanations in medicine. What we need to do is to only develop this distinction as an explanatory distinction *within* medicine.

## —3.3 Addressing the motivations behind the disease versus moral-problem-in-living distinction

The soft disease versus moral-problem-in-living distinction was motivated by certain issues, such as the stigmatisation of certain conditions, funding, the promotion of medical professionalism, and the prevention of the medicalisation of moral problems of living. I will present alternatives to how those issues can be addressed that do not depend on the distinction.

One argument for using the disease versus moral-problem-in-living distinction to constrain medicine is that the distinction helps fight the social and moral stigma of certain conditions, such as addiction (Leshner, 1997). But some have pointed to empirical evidence suggesting that the de-stigmatisation of calling a condition a disease (and therefore classifying it as a non-moral problem) is mixed at best and can itself have stigmatising consequences (Racine et al, 2015; Frank and Nagel, 2017). Some have argued that, in clinical contexts where agency is involved, an alternative solution to addressing blame (closely associated with stigma) is to separate blame from responsibility, so-called 'responsibility without blame' (Pickard, 2011; Pickard, 2017). While 'solving' the problem of stigma is beyond the scope of this paper, it is not at all obvious that calling something a disease reduces stigma. We need to consider the potential 'side effects' of medicalisation in undermining the role that choice may play in a condition (Kvaale, et al. 2013).

Concerning funding, the disease versus moral-problem-in-living distinction has helped increase funding into the research and treatment of certain conditions for conditions like addiction. This can be seen most clearly in the re-branding of addiction as a brain disease and arguments that addiction is therefore not a moral-problem-in-living (Leshner, 1997). However, such funding schemes have come with non-trivial costs. It has been argued by some scholars that the increase in funding of addiction because of the 'brain disease rebranding' of addiction has led to favouritism towards models that presume the brain disease model of addiction over other models (even though other models might have more promise) (Satel and Lilienfeld, 2014). By arguing that addiction is a disease and not a moral problem, acknowledging the role of agency can play now puts its medical status (and thereby its funding) at risk. But this is because the distinction itself accepts that, if a condition is more involuntary, then it is more worthy of funding and care. The problems of stigma and funding appear to be made worse by the disease versus moral-problem-in-living distinction, as they limit other avenues for explanation and treatment.

A better approach to addressing stigma and funding would be to argue that we should fund and not stigmatise conditions not because they are involuntary or nonmoral, but because they cause suffering for the people who have those conditions. The disease versus moral-problem-in-living distinction makes it harder to make that claim, and it distorts not only our thinking about conditions but also how we should think about them from a societal perspective. The concern about medical professionalism is that it requires some kind of disease versus moral-problem-in-living distinction. This concern is based on several claims, namely, that medicine should not be concerned with policing people's behaviours and that such moral concerns are personal, or subjective, or social rather than strictly medical in nature (Sadler, 2005 p. 224). If we reject the disease versus moral-problem-in-living distinction as a means to constrain the scope of medicine, as I argue that we should, how then is medical professionalism maintained?

This concern over medical professionalism is based on a narrow conception of morality as concerned primarily with negatively evaluating others and imposing one's own particular morality on others. This indeed should not be done and is unprofessional. But if broadly construed, morality and moral-problems-in-living concern questions of the self and how to act and need not entail shaming and blaming others. Moreover, to the extent that the disease versus moral-problem-in-living distinction is based on this narrow conception of morality, maintaining the distinction perpetuates this narrow conception of morality. The implication of the distinction may be that harshly judging and blaming others is only bad when applied to the realm of medicine, but is perfectly acceptable when it comes to problems in the realm of morality. But this implication does not follow, and perhaps such a conception of morality should be rejected anyways.

Then, there is the concern that without a disease versus moral-problem-in-living distinction that moral-problems-in-living will be medicalised. This concern can be presented in two ways. The first is applying medical approaches (narrowly construed) to moral problems of living. 'Narrowly' here refers to explanations that make minimal or no reference to questions of choice or the self. The second way concerns the problem of expanding our very conception of medicine and medicalisation to include problems of a moral nature. The disease versus moral-problem-in-living distinction is supposed to address both of these problems. This second problem, however, is just another version of the concern over medical professionalism, and is only really a problem if we accept a narrow conception of morality that should be rejected anyways.

The first problem is only a problem if we accept an overly narrow conception of medicine. That is, it is a problem to apply medical explanations and treatments to moral-problems-in-living and moral explanations and treatments to medical problems only if medicine and morality are both narrowly construed. But if medicine concerns a wider range of problems and if morality is not just concerned with negative moral judgments of actions but with positive guidance of the self, then there is no inherent conflict between medicalisation and moralisation. We can prevent the misapplication of narrower explanations to each other by developing the disease versus moral-problem-in-living distinction as a strictly explanatory distinction and not as a distinction to constrain medicine. However, that distinction has so far not been well developed. This may in part be because there is no benefit to develop such a distinction from the perspective of medicine, since falling on the moral side of that distinction might threaten the medical status of certain conditions. Here, again, we see that (theoretically at least) the disease versus moral-problem-in-living distinction to constrain the scope of medicine threatens explanatory concerns, or makes them subordinate to, issues of maintaining the scope of medicine. The bottom line, then, for why we should not use the disease versus moral-problem-in-living distinction to constrain the scope of medicine should is because doing so fundamentally assumes that we can cleanly split human problems into clearly medical versus moral problems. But such a distinction cannot be found in nature because human beings are not just their bodies nor just their selves, but fundamentally both. To attempt to separate them leads the problems mentioned above. We should therefore reject the project of constraining medicine to non-moral problems and deal with the issues that motivated this project by expanding our conceptions of medicine and morality and developing the disease versus moral-problem-in-living distinction as an explanatory distinction within medicine.

**Conclusion**

The purpose of this paper has been twofold. First, to clarify the disease versus moral-problem-in-living distinction. Second, to argue against using either soft or hard versions of the distinction to constrain the scope of medicine. Medical and moral terms have been used in many different ways to relate to one another. It has not been my purpose to provide an answer to the question of how we in all cases should distinguish moral from non-moral problems. Rather, it has been my purpose to capture how these terms have been used and how we should think about these broad categories in general. I have been primarily concerned with the purpose of rethinking the relationship between moral from non-moral problems within the context of medicine, not with how we in all cases determine the medical or moral status of particular conditions.

Medical and moral concepts have been used in a variety of ways throughout history. At certain times, such concepts have been used interchangeably to refer to the same phenomena, such as describing addiction as both a disease, vice, and a sin. At other times, some medical and moral concepts were used in an explanatory sense, such as explaining a condition as a disease as opposed to as a moral problem. However, such use of medical and moral terms were not always meant to separate the domain of medicine and morality as such. Indeed, Benjamin Rush's vision for medicine was that both vices and diseases rightly belonged under the purview of medicine. While Rush's vision was not universally subscribed to, neither was the view that we are now more familiar with, namely, that medicine should be limited primarily to the explanation and treatment of diseases in the narrow sense of the term, i.e., involuntary and non-moral conditions. Such a view developed slowly in response to different trends and concerns, such as the desire to professionalise medicine by making it more scientific and more distinct from religious and lay healers, as well as a desire to remove overtly moral and negative judgments from medicine such as blame and shame. A hard disease versus moral-problem-in-living distinction was therefore not just motivated by explanatory reasons, but also for the purpose of serving non-explanatory goals such as professionalisation. But such a view was also based on an overly narrow conception of disease and by extension of medicine, i.e., diseases as biological, involuntary, non-moral. I have argued elsewhere (see paper two) that it is this narrow conception of diseases as involuntary that motivated Szasz's arguments that medicine was medicalising problems of a voluntary and moral nature, and that therefore, for Szasz, the solution to such medicalisation was to constrain the scope of medicine to non-moral problems (Szasz, 1961). Such a view of disease was rejected later on in the debate about the meaning of the concept of disease in favour of a broader conception of disease defined by its negative consequences, instead of a particular aetiology (Kendell, 1975). While the narrow conception of disease that the hard disease versus moral-problem-in-living distinction was based on was rejected, how exactly to think about the relationship between medical and moral kinds was not clear. What replaced the hard disease versus moral-problem-in-living distinction was a soft version of the distinction that also rejected the narrow conception of disease, but that was still motivated to separate the domains of medicine and morality to prevent misapplications, secure medical professionalism, reducing stigma and blame, and gaining acceptance for the medical legitimacy of medicine and psychiatry.

The soft version of the disease versus moral-problem-in-living distinction is based on the assumption that we can separate medical disorders from certain moral-problems-in-living without undermining our commitment to a broad conception of disease. In particular, a certain kind of 'compatibilism' between medical disorders and moral problems of living has been argued for by some, drawing on the work of John Sadler to make a soft disease versus moral-problem-in-living distinction (Zachar and Potter, 2010a; Zachar and Potter, 2010b; Zachar, 2011). Such a position is based on the view that, as long as a condition meets the general criterion for being a medical disorder, then a condition can be both a medical disorder and have significant moral components (Sadler, 2005). The underlying assumption here is that such a medical criterion provides, or should provide, indispensable non-moral features that prevents the medicalisation of moral-problems-in-living. And the underlying motivation is still that medical disorders and moral-problems-in-living should be separated as much as possible, and that those medical disorders with moral components should have such components translated as much as possible into non-moral language. Such a motivation may be in part be to simply remove moral language that reflects particular 'societally-popular' views, such as views about sexual morality, that should be removed from our moral thinking in general. But such a concern appears to also be motivated by a desire to make sure that medicine does not expand into the realm of moral problems of living, in the same way that motivated Szasz to a much greater extent than the proponents of the soft medicine versus morality distinction: to remove from medicine those moral-problems-in-living that are rightly moralised because they *are* moral-problems-in-living. I have argued in this paper that any attempt to establish a soft disease versus moral-problem-in-living distinction to constrain the scope fo medicine necessarily undermines our commitment to a broad conception of disease defined by their negative consequences and not by a particular aetiology. Dysfunction cannot serve as a non-moral criterion to distinguish genuine medical disorders from moral problems of living because dysfunctions can be moral in nature. The only way to distinguish moral from non-moral dysfunctions that has been proposed is to consider the role that agency plays in moral versus non-moral dysfunctions. But whatever that role may turn out to be, introducing any kind of 'involuntariness' necessitates an (unwanted) return to a narrower conception of disease.

We should therefore not constrain the scope fo medicine to non-moral problems via a soft disease versus moral-problem-in-living distinction. Not only because it is incompatible with the shared commitment to a broad conception of diseases and by extension of medicine, but also because any commitment to such a distinction necessarily introduces incentives against voluntary and moral explanations in medicine and incentives in favour of involuntary and non-moral explanations. Such incentives, in turn, create a potential conflict between a condition's medical status and the development of the best explanations (and thereby treatments) for that condition. Szasz wanted to prevent the medicalisation of moral-problems-in-living by using the disease versus moral-problem-in-living distinction to constrain the scope of medicine. Others want such a distinction, whether hard or soft, to prevent the moralisation of medical problems. I have argued that medicalisation poses a threat to agency and moral problems of living precisely when we eject them from the domain of medicine, and that the best way to prevent the misapplication of medical and moral concepts is to better develop such a distinction *within* a wider conception of medicine.

**References**


Agich, G. J. 1994. Evaluative judgments and personality disorder. *Philosophical Perspectives on Psychiatric Diagnostic Classification*. J. Z. Sadler, O. P. Wiggins, and M. A. Schwartz, eds. Baltimore, MD: The John Hopkins University Press: 233-245.

Banicki, C. 2018. Personality Disorders and Thick Concepts. *Philosophy, Psychiatry, & Psychology,* 25,3: p. 209-221.

Charland, L.C. 2004. Character: Moral treatment and the personality disorders. In The philosophy of psychiatry: A companion, ed. J. Radden, 64–78. Oxford: Oxford University Press.

Charland, L.C. 2006. The moral character of the DSM IV cluster B personality disorders. Journal of Personality Disorders 20(2): 116–125.

Charland, L. C. 2007. Benevolent theory: moral treatment at the York Retreat. *History of Psychiatry*, 18:1, p. 61-80.

Charland, L. C. 2008. A moral line in the sand: Alexander Crichton and Philippe Pinel on the psychopathology of the passions. In L. C. Charland & P. Zachar (Eds.), *Fact and value in emotion*. Amsterdam, The Netherlands: John Benjamin Press.

Charland, L.C. 2010. Medical or moral kinds? Moving beyond a false dichotomy. Philosophy, Psychiatry, & Psychology 17(2): 119–125.

Charland, L. C. 2011. Moral undertow and the passions: two challenges for contemporary emotion regulation. *Emotion Review*, 3:1, p. 83-91.

Chavigny, K. A. 2013. “An Army of Reformed Drunkards and Clergymen”: The Medicalization of Habitual Drunkenness, 1857–1910. *Journal of the History of Medicine and Allied Sciences* 69:3, p. 383-425.

Crothers, TD. 1891. ARE INEBRIATES CURABLE? *JAMA.* 1891;XVII(24):923-927. doi:10.1001/jama.1891.02411020021001a

Downame, J. 1609. Foure Treatises Tending to Diswade All Christians from Foure no Less Hainous then Common Sinnes: Namely, the Abuses of Swearing, Drunkennesse, Whoredome, and Briberie. London: Felix Kyngston.

Ferentzy, P. 2001. From sin to disease: differences and similarities between past and current conceptions of chronic drunkenness. Contemporary Drug Problems 28 (Fall), p. 363-390.

Fingarette, Il. 1988. Heavy drinking: The myth of alcoholism as a disease. University of California Press.

Foucault, M. 1973. The Birth of the Clinic: an Archaeology of Medical Perception. London: Tavistock.

Frank, L. E. & Nagel, S. K. 2017. Addiction and Moralization: the Role of the Underlying Model of Addiction. *Neuroethics* 10: p. 129–139.

Heather, N. 2013. Is alcohol addiction usefully called a disease? *Philosophy, Psychology, & Psychiatry* 20:4, p. 321-324.

Hershon, H. 1974. Alcoholism and the Concept of Disease. *British Journal of addiction* 69, p. 123-131.

Heyman, G. M. 1996. Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19, p. 561-610.

Heyman, G. M. 2009. Addiction: A Disorder of Choice. Cambridge, MA, Harvard University Press.

Heyman, G. M. 2013. Addiction and choice: theory and new data. *Frontiers in psychiatry: Perspective Article,* 4:31, 1-5.

Heyman, G. M. 2017. Do addicts have free will? An empirical approach to a vexing question. Addictive Behaviors Reports 5, 85–93.

Horne, G. 2014. Is Borderline Personality Disorder a Moral or Clinical Condition? Assessing Charland's Argument from Treatment. *Neuroethics* 7:215–226.

Jellinek, E. M. 1946. Phases in the drinking history of alcoholics. *Quarterly Journal Studies on Alcohol* 7, p. 1-88.

Jellinek, E. M. 1952. Phases of alcohol addiction. *Quarterly Journal of Studies on Alcohol* 13, p. 673-84.

Jellinek, E. M. 1960. The disease concept of alcoholism. Highland Park, NJ; Hillhouse.

Johnstone, G. 1996. From vice to disease: The concepts of dipsomania and inebriety, 1860-1908. Social Legal Studies, 5(1), 37-56.

Keller, M. 1976. The disease concept of alcoholism revisited. Journal of studies in Alcoholism 37: p. 1694-1717.

Kennett, J. 2013. Addiction, Choice, and Disease: How Voluntary Is Voluntary Action in Addiction? In Neuroscience and Legal Responsibility, edited by Nicole A. Vincent, OUP. pp. 258-278.

Kvaale, E.P., N. Haslam, and W.H. Gottdiener. 2013. The side effects of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. Clinical Psychology Review 33: p. 782–794.

Leshner, A. I. 1997. Addiction Is a Brain Disease, and It Matters *Science* 278, p. 45-47. DOI: 10.1126/science.278.5335.45

Leshner, A.I. 2001. Addiction is a brain disease. Issues in Science and Technology Online, 17(3). http://issues.org/17-3/leshner/. Accessed 23/05/2016.

Levine, H.G. 1978. The discovery of addiction: Changing conceptions of habitual drunkenness in America. J. Stud. Alcohol 39: 143-174, 1978.

Lewis, D. C. 1991. Comparison of Alcoholism and Other Medical Diseases: An Internist's View. *Psychiatric Annals* 21,5; p. 256-265.

Lewis, M. 2015. The biology of desire: why addiction is not a disease. New York: Public Affairs.

Lewis, M. 2017. Addiction and the brain: Development, not disease. *Neuroethics* 10, p. 7-18.

Lewis, M. 2018. Brain Change in Addiction as Learning, Not Disease. *The New England Journal of Medicine* 379;16, p. 1551-1560.

Maltzman, I. 1991. Is Alcoholism a Disease? *A Critical Review of a Controversy. Integrative Physiological and Behavioral Science*, 26,3, p. 200-210.

Martin. M. W. 2010. Personality Disorders and Moral Responsibility. Philosophy, Psychiatry, & Psychology 17(2): 127–129.

Massing, M. 2000. Seeing drugs as a choice or as a brain anomaly. *New York Times*. Retrieved from <www.nytimes.com> on 14-03-2014.

Miller, N. S. and Chappel, J. N. 1991. History of the Disease Concept. *Psychiatric Annals* 21,4; 196-205. Edit.

Nash, J. O. 1894. Review of "Drunkenness, by George R. Wilson. 161 pp. Crown 8vo. 2.s 6d. sonnenshein. London 1893." In The Economic review, 4,2: p. 289-290

Pearce, S. & Pickard, H. 2010. Finding the will to recover: philosophical perspectives on agency and the sick role. *Journal of Medical Ethics.* Online. 10.1136/jme.2010.035865

Pearce, S. 2011. Answering the Neo-Szaszian Critique: Are Cluster B Personality Disorders Really So Different? *Philosophy, Psychiatry, & Psychology,* 18:3, p. 203-208.

Pickard, H. 2011. What is personality disorder? *Philosophy, Psychiatry, & Psychology*, 18:3, p. 181-184.

Pickard, H. 2011. Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder. P*hilosophy, Psychiatry, & Psychology,* 18:3, p. 209-224.

Pickard, H. 2017. Responsibility without Blame for Addiction. *Neuroethics* 10: p. 169–180.

Pies, R. 1979. On myths and countermyths. *Archives General Psychiatry* 36:2, p. 139-144.

Potter, N. N. 2013. Moral evaluations and the cluster B personality disorders. *Philosophy, Psychiatry, & Psychology, 20,* 217–9.

Racine, E., Bell, E., Zizzo, N., & Green, C. 2015. Public Discourse on the Biology of Alcohol Addiction: Implications for Stigma, Self-Control, Essentialism, and Coercive Policies in Pregnancy. *Neuroethics* 8: p. 177–186.

Reimer, M. 2013. Moral disorder in the DSM-IV? The cluster b personality disorders. Philosophy, Psychiatry and Psychology, 20, 203–15.

Reimer, M., and B. Day. 2013. Affective Dysfunction and the Cluster B Personality Disorders. *Philosophy, Psychiatry, & Psychology,* 20,3, p. 225-229.

Rush, B. 1805. An Inquiry into the Effects of Ardent Spirits upon the Human Body and Mind, 4th edition. Philadelphia: Printed for Thomas Dobson, 1805.

Rush, B. 1812. Medical Inquiries and Observations upon the Diseases of the Mind. Philadelphia: Kimber & Richardson.

Sadler, J. Z. 2005. *Values and Psychiatric Diagnosis*. Oxford: Oxford University Press.

Sadler, J. Z. 2008. Vice and the Diagnostic Classification of Mental Disorders: A Philosophical Case Conference. *PPP* 15,1, p. 1-17.

Sadler, J. Z. 2013. Values in Psychiatric Diagnosis and Classification. In The Oxford Handbook of Philosophy and Psychiatry Edited by K.W.M. Fulford, Martin Davies, Richard G.T. Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton. Oxford: Oxford University Press. p.

Satel, S. and Goodwin, F. 1997. Is Addiction a Brain Disease? *Ethics and Public Policy.* Available from: www.eppc.org/docLib/20030420_ DrugAddictionBrainDisease.pdf

Satel, S., & Lilienfeld, S. O. 2014. Addiction and the brain-disease fallacy. *Frontiers in Psychiatry*: Review Article, 4:141, p. 1-11.

Scadding, J. G. 1967. Diagnosis: the clinician and the computer. *Lancet* 2, p. 877-882.

Schuur, R. 2019. Mental Health and Illness: Past Debates and Future Directions. Bloomsbury Companion to Philosophy of Psychiatry, edited by S. Tekin and R. Bluhm. p. 527-541.

Segal, G. M. A. 2013. Alcoholism, disease and insanity. *Philosophy, Psychiatry, & Psychology* 20,4: p. 297–315.

Spitzer, R. L. & Endicott, J. 1978. Medical and mental disorder: Proposed definition and criteria. In: Spitzer, RL.; Klein, DF., editors. Critical Issues in Psychiatric Diagnosis. New York, NY: Raven Press; p. 15-39.

Southworth, J. 2013. Can Morally Disvalued Traits Constitute the Symptoms of a Mental Disorder? *Philosophy, Psychiatry, & Psychology,* 20,3, p. 221-223.

Szasz, T. S. 1961. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins.

Szasz, T. S. 1974. The Myth of Mental Illness: Foundations of A Theory of Personal Conduct. New York: Harper Collins.

Szasz, T. S. 1972, 'Bad habits are not diseases: A refutation of the claim that alcoholism is a disease', *The Lancet 2,* 83—84.

Todd, J. E. 1883. Drunkenness a vice, not a disease. Hartford; Case, Lockwood & Brainard.

Vaillant, G. E. & Milofsky, E. S. 1982a. The etiology of alcoholism. *American Psychologist* 37:494-503.

Vaillant, G. E. & Milofsky, E. S. 1982b. Natural history of male alcoholism: 4. Paths to recovery. *Archives of* General Psychiatry 39:127-33.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundaries between biological facts and social values. *American Psychologist* 47:3, p. 373–388.

Wakefield, J. C. 2017. Addiction and the Concept of Disorder, Part 2: Is every Mental Disorder a Brain Disorder? *Neuroethics* 10, p. 55-67.

Warner, J. 1994. Resolv'd to drink no more: Addiction as a preindustrial construct. Journal of Studies on Alcohol 55, p. 685-691.

Williams, B. 1986. Ethics and the Limits of Philosophy. Harvard University Press: Cambridge, MA.

Zachar, P., and N. N. Potter. 2010a. Personality disorders: moral or medical kinds—or both? Philosophy Psychiatry & Psychology 17(2): 107–117.

Zachar, P., and N. N. Potter. 2010b. Valid Moral Appraisals and Valid Personality Disorders. Philosophy Psychiatry & Psychology 17(2): 131–142.

Zachar, P. 2011. The clinical nature of personality disorders: Answering the neo-Szazian critique. Philosophy, Psychiatry, & Psychology, 18,3, 191–202.