The ethics of involuntary psychiatric treatment: A study into the social and moral structures through which we construct illness and the justifications for involuntary intervention

Craig Edwards
B. Law, B. Comm., Dip Arts (Hons).

This thesis is presented for the degree of Doctor of Philosophy
of The University of Western Australia

School of Humanities
Discipline of Philosophy

2012

Dissertation Abstract

This dissertation takes the form of a series of papers which, if taken together, study the role of patient autonomy in involuntary psychiatric treatment. This is not an argument 'for' or 'against' involuntary psychiatric treatment. My reasoning supports the status quo insomuch as involuntary hospitalisation and treatment are sometimes warranted, with the greater moral question being that of *when* is involuntary treatment justified. That is, what moral framework should guide institutional policies on the use of involuntary treatment? I argue that the orthodox moral theory fails to adequately account for both the evaluative component of mental competence and the social and relational features of meaningful personal autonomy. These deficiencies in the underlying moral framework have forced psychiatrists and medical institutions to continually stretch formal criteria for intervention to the point where they have become ad hoc justifications for paternalism, in which bias can be embedded in a manner that is undetectable in individual cases, and impervious to effective challenge by individual patients. Following the writings of O'Neill and other philosophers concerned with the autonomy of patients and similarly disempowered participants in institutional settings, I argue that patient autonomy arises not from independence, but from morally health relationships of care and trust. In the first half of this dissertation, I endeavour to construct such an account of patient autonomy as it relates to involuntary psychiatric treatment, firstly through exploring the normative qualities of illness as a moral concept, reviewing the liberal tradition in which medical ethics operates, and then building upon this an account of patient autonomy where autonomy is achieved through authenticity in the pursuit of the patient's goals and values, rather than a right to qualify for independent choice. In the second half of this dissertation, I apply this to a series of applied problems within the ethics of involuntary treatment, both as worthwhile topics of inquiry in their own right and as means of further developing this

account of patient autonomy.

Table of Contents

Acknowledgments:

The following people have provided supervision during the course of this dissertation, though special acknowledgment should be made of the direct supervision provided by Professor David Van Mill for the majority of my dissertation, and Dr Keith Horton for the first year of my dissertation:

Prof. David Van Mill

Dr Keith Horton

Prof. Michael Levine

Prof. Stewart Candlish

I also thank John Kleinig for making himself available for consultation regarding the topic of paternalism during the initial phase of my research.

Candidate Declaration

This thesis **does not contain** work that I have published, nor work under review for publication.

Introduction

1. Overall aims

This dissertation takes the form of a series of papers which study the role of patient autonomy in involuntary psychiatric treatment. I intend that these articles can stand alone as independent works, but also go towards the formation and illustration of a broader account of autonomy and psychiatric paternalism. Towards this end, I have created two adjoining sets of works, forming Part A and B of this dissertation. The papers that comprise Part A are concerned with questions of moral theory and conceptual analysis. They provide an initial account of patient autonomy and its relationship to psychiatric paternalism. The papers that form Part B are more practical, each examining a separate question of applied psychiatric ethics, while expanding upon the basic account established in Part A. Through these works as a whole I defend my thesis, which is that the existing bioethical orthodoxy misconceives patient autonomy in a way that corrupts our moral reasoning with regard to the use of involuntary treatment.

This is not an argument 'for' or 'against' involuntary psychiatric treatment. The tragic reality of mental illness is that *some* patients, in *some* situations, face such terrible harm that involuntary treatment is sadly warranted. Blanket opposition to involuntary psychiatric treatment has been pushed to the very fringes of philosophy, and is almost entirely drawn from the arguments (often more philosophical than medical) of psychiatric skeptics such as Thomas Szasz (1974; 1998) whose remain infamous yet have become largely irrelevant. However, it would be unfortunate if our moral acceptance of involuntary illness lulled us into moral complacency. Involuntary psychiatric treatment remains a dangerous inversion of the ordinary doctor-patient relationship, carrying both the practical risk of abuse through a heavily imbalanced

power arrangement, and the moral hazard of eroding the patient's entitlements to autonomy and respect. We should approach involuntary treatment as a legitimate tool, but a dangerous one, to be viewed with vigorous skepticism. My thesis, then, is to ask what moral framework should guide institutional policies on the use of involuntary treatment. I argue that the orthodox moral theory fails to adequately account for both the evaluative component of mental competence and the social and relational features of meaningful personal autonomy. These deficiencies in the underlying moral framework have forced psychiatrists and medical institutions to continually stretch formal criteria for intervention to the point where they have become ad hoc justifications for paternalism, in which bias can be embedded in a manner that is undetectable in individual cases, and impervious to effective challenge by individual patients.

Following the writings of O'Neill and other philosophers concerned with the autonomy of patients and similarly disempowered participants in institutional settings (Bluhm 2009; Oshana 2003; Oshana 2005; O'Neill 2002; O'Neill 2003; Meyers 1989; Meyers 2005), I argue that the most fundamental components of patient autonomy are not those relating to independence, but those that enable morally healthy relationships of care and trust. The moral status of paternalistic intervention should be judged by the values that characterise the resulting treatment relationship. Instead of setting a standard of mental competence that patients must pass to qualify for control of their medical choices, medical staff respect a patient's autonomy if treatment choices (including the decision whether to impose involuntary treatment) are made in accordance with that patient's own values, protecting the patient from having the values of other individuals or institutions imposed upon her – just as they are tasked with protecting the patient from imposition by mental illness upon her character,

A doctor faced with immediate care of a series of potentially ill patients cannot be expected to take account of the great variety of philosophical and legal debate relevant to that choice, and it is a misinterpretation of the philosophers' aim to view the philosophy of bioethics in that context. Realistically, the doctor – following her professional training and legal advice – should turn to her rules for professional practice and hopefully find a brief and clear set of considerations that she can apply even with limited knowledge and time. However, as the doctor progresses in her profession, a time may come when it is no longer enough to justify her decisions by relevance solely to these brief and clear considerations. She may begin to ask about the goals that these considerations should ultimately be guided by. She may have noticed over the years that cases have arisen where she has applied the stipulated considerations fully, yet feels like justice was somehow lacking. She may even seek to suggest improvements to the considerations and, in the process, find a need to reconcile competing moral interests.

The former kind of consideration, i.e. the brief and clear set considerations that the doctor applies on a day to day basis, is best dealt with by those close to the process it affects: doctors, patients, and patient advocates rather than academic philosophers. Many philosophers may disagree with me on that, but we can put such differences aside as irrelevant to the work at hand. Rather, I am more interested in the kind of activity that follows the doctor's attempts to develop an improved professional practice rule, and the much broader deliberation that results. My aim is to illuminate the concepts that should inform such efforts, to provide a moral framework against which psychiatry's professional and legal codes can be judged, and to establish a set of moral principles that they should strive to achieve.

2. Context

This is a work *of* liberalism with regard to medical paternalism, but not strictly speaking an argument in demonstration of liberalism. I am not constructing a comprehensive proof that medical or psychiatric treatment is part of the patient's personal domain, within which her own interests and personal autonomy warrants the kind of limitations upon autonomy suggested by the liberal philosophical traditions. My justification for this is entirely practical: it is a long-held legal and professional rule of conduct that a patient's informed consent must be obtained before any medical procedure (Kirby 1983; Annis 1984), and although I illustrate flaws in the traditional conception of this principle, medical treatment is near-universally considered a crucial part of one's personal domain, subject to liberal restrictions upon autonomy. A justification of medical liberalism, from the ground up, would easily occupy a book, much of which would be spent arguing points that are already broadly accepted. Even those works that call for radical reformation start from a perspective of medical liberalism and work to redefine patient autonomy in arguing for practical reform (O'Neill 2002; Radden 2004; Stirrat and Gill 2005). Instead of arguing *for* liberalism on paternalism, I position this philosophical work *within* the liberal tradition. I aim to demonstrate what it is we commit to in adopting respect for the agency of other persons, and to examine the scope of paternalism that remains permissible despite – and sometimes because of – this respect.

Liberalism with regard to paternalism is a *very* different philosophical standpoint to political liberalism: it has nothing whatsoever to do with the valuing of individual freedom as against the claims of society and its institutions. At a prima facie level, liberalism as a position on paternalism is compatible with any broader political

morality from libertarianism to socialism. It is an entirely personal standpoint, addressing the matter of autonomy as it applies to the one person's *own* interests. To be strictly liberal here is to prioritise one person's autonomy over *that same person's* well-being. As soon as some other party's interests enter the moral question, the relevant consideration is no longer one of liberalism in this special sense – that of a liberal approach to paternalism.

Misguided paternalism is not the greatest problem facing psychiatric patients today. Yet it remains an important one, and I fear that in recent years, a heightened interest in patient well-being and concern about diminished care and resource shortages following the shift away from long-term institutionalisation (Forchuk 2008; Gostin 2007; Kress 2006), may have lessened the public and academic interest in critiquing involuntary treatment. Wrongful paternalism denies people full and proper participation as members of society. To adopt the commonly used term from western liberalism, it denies them 'respect'. There is a symbolic aspect to this, where people risk being labeled 'the other' through the imposition of processes that are not required of others in society. But denial of respect is not merely an emotional insult, or an infliction of hurt feelings. It is a practical and substantial exclusion of the freedom, self-governance and participation that we are ordinarily granted as members of society. Again, I am not writing in opposition to psychiatric paternalism, whether in principle or as an institutionalised practice. But its institutional and coercive qualities make patients tremendously vulnerable, and their protection from wrongful application of psychiatric paternalism depends upon the ability of psychiatric institutions and review bodies to understand and submit to the considerations relevant to paternalism's proper implementation. If I am correct in believing that the current concept of mental competence invites moral confusion, then its ability to protect patients by testing the

proper implementation of paternalism is undermined in a manner that risks serious injustice.

3. Synopsis

This dissertation is divided into two parts and seven papers, not including this introduction and a concluding summary of findings. Three of these papers have been published in peer reviewed journals (listed below). I leave these papers largely unchanged from their original published form, however some alterations are necessary by virtue of the format. Firstly, I have altered referencing styles to maintain a consistent citation format in this dissertation. Secondly, I have occasionally omitted sections where the demands of publication as an unknown doctoral candidate required me to restate in detail theories that I have already discussed at length in earlier sections of this dissertation. I do not omit all such repetition – for example, I retain some discussion of personal integrity in 'Respect for Other Persons (Edwards 2010) because it provides a useful and brief reminder of those vital concepts at a point where they have not been described in such a form for several papers. Elsewhere, however, I have made omissions for the sake of avoiding unnecessary repetition. In order to keep these sections in a form where I describe them with honesty as the published works listed in this section, I have avoided rewording or adding any material, except where necessary for grammatical continuity following an omission. This means that due to the original publisher's in-house stylistic requirements, there will be some stylistic changes between papers.

I include a fourth published piece as a schedule to this dissertation (Edwards 2009a). In the publication of the first paper in this dissertation, 'Ethical decisions in the

classification of mental conditions as mental illnesses' (Edwards 2009b), I had the good fortune of having commentary articles follow my paper, together with a response to those commentaries by myself. To my mind, this response clarifies the logic of my original argument, and resolves an important ambiguity (I address this ambiguity in the abstract that introduces the paper in this dissertation). It may be of interest if the reader wishes further elucidation of my reasons for finding Wakefield's (1992; 2006; 2009) account of mental illness to be inadequate. Nonetheless, it does not alter or significantly add to my original account, and so I do not include it in the main body of this dissertation.

Part A: Psychiatric paternalism and the liberal tradition of medical ethics

Part A establishes the conceptual and moral framework through which I examine the ethics of involuntary psychiatric treatment. Its three component papers adopt a modular structure, addressing the concept of mental illness, the philosophical traditions underlying restrictions upon medical paternalism, and the application of those philosophical traditions in the context of mental impairment, respectively. In the first, 'Ethical decisions in the classification of mental conditions as mental illnesses' (Edwards 2009b) I argue that the identification of mental illness is a normative exercise, reflecting our conception of the proper function of personhood. It is routine to acknowledge a role for normativity in distinguishing between harmful and harmless types of mental dysfunction, but until recently dysfunction itself has been understood in purely evolutionary terms. I reason that evolutionary dysfunction is only tenuously related to mental illness, and that in our attribution of illness we determine the 'proper function' of physical and mental processes by the purposes we impose upon them. Our classification of mental illness reflects an evaluative judgment to define 'proper function' in terms of a liberal model of personhood, in which we bear broad (but

imperfect) moral autonomy and responsibility over who we are.

In the second piece, 'Reasons, autonomy and paternalism', I examine the liberal traditions that underlie current legal and ethical restrictions upon medical paternalism. Philosophers arguing for limits upon paternalism by the state have most commonly sought to cite J.S. Mill's *On Liberty* (1977, originally 1859) as a basis for their account of personal autonomy. Many of the more substantial works on paternalism are skeptical of the potential for a Kantian basis for legal rights against paternalism, citing Kant's concern with 'rational' autonomy as providing a narrower liberty than Mill's appeal to the moral and practical importance of freedom. I argue that the Kantian tradition of respect for rational autonomy is more defensible than the neo-Millian account and, if properly constructed, provides a more robust personal autonomy than its critics acknowledge. In any event, for personal autonomy to have the overriding value ascribed to it by liberal accounts of paternalism, it must take the form of an appeal to authenticity or *personal integrity*, rather than freedom of immediate choice. That is, an appeal to the goals and values that define the person's authentic self, in recognition of her interest in developing and pursuing her own conception of the good life.

The third piece, 'Mental competence and its limitations' is much longer than the other papers in this dissertation, and is itself a modular work in which I make the case for fundamental reform of the ethical framework by which the suitability of psychiatric paternalism is assessed. The informed consent doctrine, under which treatment must not be applied to a mentally competent patient without her consent, purports to provide an objective basis for judging the appropriateness of paternalistic treatment that is independent of the content of the patient's choice – i.e. that the patient is free to make

whatever treatment choices she wants, so long as she is sufficiently mentally capable in relation to each decision. By that measure, the doctrine fails in a manner that is deeply at odds with the concept of personal autonomy discussed in the previous paper; mental incompetence is routinely inferred from the unreasonableness of a choice, illustrated by the asymmetrical standards of competence required for refusal and acceptance of recommended treatment. With no formal stipulation of the values by which reasonableness is determined, the informed consent doctrine facilitates the imposition of the values held by psychiatric staff and institutions over those of the patient. Drawing upon my account of mental illness from the first paper, I then argue that mental competence is necessarily an evaluative concept, reflecting one's judgment as to the kind of individual whose personhood should be socially authenticated. In the latter part of this piece, and following from my discussion of personal autonomy in the second paper, I argue that the significance of mental competence to patient autonomy has been overstated. Competence is relevant to determining the authenticity of an individual's current wants, but a patient's autonomy is primarily a matter of authenticity in the goals and values that guide the patient's treatment. I suggest an alternative framework for assessing psychiatric paternalism, centering upon personal identity and integrity, rather than mental competence. Respect for autonomy requires that, where possible, medical staff seek to identify and apply the goals and values relevant to treatment that are authentic qualities of the patient's character, rather than the result of mental illness or coercion. Psychiatric paternalism shares the same moral basis, serving as a means of protecting our central goals and values from contrary normative impositions.

My work in Part A seeks to shift the focus of moral attention in judging psychiatric paternalism away from the search for objective standards of mental illness and

incompetence, to critique of the goals and values by which the attribution of illness, incompetence and treatment choices are governed. As a whole, it shows that justice in involuntary treatment can only follow from a just conception of *who* a person is. That is, justice in the assessment of involuntary treatment turns upon the extent of variation in mental qualities that one can choose and still have one's personhood socially authenticated (as determined by our standards of mental illness and incompetence), and our understanding of the qualities that define that person.

Part B: Applied problems in psychiatric ethics

In these papers I explore the implications of the account developed in Part A through a series of ethical puzzles arising from involuntary psychiatric treatment. In addition to addressing these applied concerns, each paper further develops the relevant moral framework, and in particular the concepts of personal identity and authenticity as they relate to respect for autonomy.

'Suicide prevention and the limits of patient autonomy' investigates the liberal model of personhood that shapes our concepts of mental illness and incompetence, and asks whether the freedom of self-authorship that liberal personhood confers is unfettered. I do this in the context of the apparent contradiction between the general restrictions upon psychiatric paternalism and policies demanding broad psychiatric intervention to prevent suicide. The evaluative nature of mental competence renders the orthodox bioethical framework incapable of meaningfully addressing this type of moral question and, similarly, my account of mental illness in Part A shows that we cannot simply cite the expansive concept of depressive illness as a justification for intervention. I argue that the liberal self does not imply a right to non-interference, but requires the positive freedoms necessary for the attribution and pursuit of value. As such, meaningful

autonomy requires a respect for one's own self-worth. This is not necessarily inconsistent with suicide: in recent years, many bioethicists have sought to defend a right to 'die with dignity' where 'dignity', in my view, is best understood as a special form of self-affirmation. Nonetheless, this construction of liberal personhood supports the use of involuntary psychiatric treatment to prevent and treat suicidality (i.e. a powerful mental disposition towards suicide), where the suicidality takes the form of an unreasonable denial of self-worth.

In 'Beyond mental competence' (Edwards 2010), I set out my account of personal identity over time, addressing the question of how the inauthentic changes to character and judgment imposed by mental illness are distinguished from ordinary personal development. This fills out the account of personal integrity that I develop in Part A, but also concerns an often overlooked problem that I call the issue of 'judgment shift'. Judgment shift is when a person's character is altered in such a manner that it seems to render her decisions on a topic deeply inauthentic, even though she would (post-change) wholeheartedly endorse the consequences of those choices as an expression of her deep-held goals and values. I seek to explain our sense of inauthenticity as a justified response to an illness-induced alteration in personal identity. Although I cite Schechtman (1996, 94-114), DeGrazia (2005, 136-170) and similar accounts as my starting point in establishing what it is to have an authentic personal identity, I dispute their common position that we should strictly distinguish questions about the authenticity of our character from those about the requirements for personal continuity. Rather, I argue that character, rightly characterised by those authors as essentially narrative in structure, is both temporally extensive and severable. In the event of judgment shift, mental illness alters a person's character in a manner that is narratively incomprehensible, radically separating her from the relationships and connative states

that characterise her authentic self.

In ‘Respect for other selves’ (Edwards 2011), I continue the account of personal identity given in the previous paper, and extend it to the matter of how treatment should be governed following permanent mental incapacity. There is a broad consensus within bioethics that such treatment should be governed by the goals and values that best characterised the patient’s self prior to her mental incapacitation, and this has led to much philosophical support for the authority of advance directives stipulating the limits of future treatment. However, a minority of philosophers have retained concerns that to give advance directives unquestioned authority risks injustice where the mentally impaired patient’s *current* experiential desires give her an interest in continued life-sustaining treatment beyond that which is permitted by her advance directive. I believe that most philosophers have underestimated the merit in this concern, and having made personal identity central to my own account of ethical imposition of treatment, this conflict between past and current interests is pertinent to determining the extent to which a patient’s pre-impairment ‘authentic’ identity should be prioritised over her other interests. I argue that, following our recognition with regard to non-human animals that personal agency is no pre-requisite for serious moral worth, we cannot so easily dismiss the patient’s ongoing interest in continuing a happy and contented way of life – just as the patient’s prior mental sophistication would not give her the moral authority to demand that her later mentally impaired self be tortured or demeaned (say, in the continuation of a rare religious practice, or an ideological opposition to the use of painkillers).

I alter the lines upon which the debate should been conducted. Previously, such issues have too often been viewed as puzzles of personal identity, in which the continuity of

the pre-impairment person trumps the discordant moments of engagement, emotion and experiential pleasure that forms the life of the post-impairment patient. I view the scenario as one of reconciling two *mutually legitimate*, but competing, moral interests. Instead of immediately assuming the authority of the pre-impairment person through her advance directive, we must examine the extent to which extending her life will undermine the projects that gave her existence meaning, just as we must examine the well-being, happiness and prospects for ongoing experiential happiness from a continued life.

In the final paper, 'Problems with the easy justification', I turn my attention to involuntary psychiatric treatment for the purpose of protecting others, rather than the paternalistic assistance of the person being treated. In contrast with the often controversial issue of paternalistic treatment, there has been hardly any critical examination of this justification for psychiatric intervention, despite an enormous amount of contemporary criticism of detention of other social groups (previously convicted sex offenders and terrorism suspects) for the same purpose. I present a case for rejecting non-paternalistic involuntary psychiatric treatment, as a gross injustice similar in nature to racism and sexism. I also suggest that the prevention of illness-induced violence can more justifiably be addressed as a matter of psychiatric paternalism, under the account developed earlier in this dissertation.

4. Publications arising from this disssertation

Edwards, 'Ethical Decisions in the Classification of Mental Conditions as Mental Illness', *Philosophy, Psychiatry and Psychology* (2009) 16,1: 73-90

Edwards, ‘Changing Functions, Moral Responsibility, and Mental Illness’, *Philosophy, Psychiatry and Psychology*, (2009) 16,1: 105-107

Edwards, 'Beyond Mental Competence', *Journal of Applied Philosophy* (2010) 23,3: 278-289

Edwards, ‘Respect for other selves’, *Kennedy Institute of Ethics* (2011) 21,4: 349-378

References (Introduction)

Annis, David B. 1984. 'Informed Consent, Autonomy, and the Law'. *Philosophy Research Archives* 10: 249-259.

Bluhm, Robyn. 2009. 'Evidence-based Medicine and Patient Autonomy'. *International Journal of Feminist Approaches to Bioethics* 2 (2): 134-151.

DeGrazia, David. 2005. *Human Identity and Bioethics*. New York: Cambridge University Press.

Edwards, Craig. 2009a. 'Changing Functions, Moral Responsibility, and Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 105-107.

———. 2009b. 'Ethical Decisions in the Classification of Mental Conditions as Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 73-90.

———. 2010. 'Beyond Mental Competence'. *Journal of Applied Philosophy* 27 (3): 273-289. doi:10.1111/j.1468-5930.2010.00491.x.

———. 2011. 'Respect for Other Selves'. *Kennedy Institute of Ethics Journal* 21 (4): 349-378.

Forchuk, Cheryl. 2008. 'Some Psychiatric Survivors Can't Survive the System'. *The Canadian Nurse* 104 (8): 44.

Gostin, Lawrence. 2007. '"Old" and "New" Institutions for Persons with Mental Illness: Treatment, Punishment or Preventative Confinement?' *Public Health* 122 (9): 906-913.

Kirby, Michael. 1983. 'Informed Consent: What Does It Mean?' *Journal of Medical Ethics* 9: 69-75.

Kress, Ken. 2006. 'Rotting with Their Rights On: Why the Criteria for Ending Commitment or Restraint of Liberty Need Not Be the Same as the Criteria for Initiating Commitment or Restraint of Liberty, and How the Restraint May Sometimes

Justifiably Continue After Its Prerequisites Are No Longer Satisfied'. *Behavioral Sciences and the Law* 24: 573-598.

Meyers, Diana. 1989. *Self, Society and Personal Choice*. New York: Columbia University Press.

———. 2005. 'Five Faces of Selfhood'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 27-55. Cambridge: Cambridge University Press.

Mill, John Stuart. 1977. *On Liberty.* Ed. John Robson. The Collected Works of John Stuart Mill. Toronto: University of Toronto Press. http://oll.libertyfund.org/title/233.

O'Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

———. 2003. 'Some Limits of Informed Consent'. *Journal of Medical Ethics* 29: 4-7.

Oshana, Marina. 2003. 'How Much Should We Value Autonomy?' *Social Philosophy and Policy* 20 (2): 99-126.

———. 2005. 'Autonomy and Self-Identity'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 77-97. Cambridge: Cambridge University Press.

Radden, Jennifer. 2004. *The Philosophy of Psychiatry: A Companion*. New York: Oxford University Press.

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornwell University Press.

Stirrat, Gordon, and Robin Gill. 2005. 'Autonomy in Medical Ethics After O'Neill'. *Journal of Medical Ethics* 31 (3) (March): 127-130. doi:10.1136/jme.2004.008292.

Szasz, Thomas. 1974. *The Myth of Mental Illness*. 2nd ed. New York: Harper and Row.

———. 1998. 'Commentary on "Aristotle's Function Argument and the Concept of

Mental Illness"'. *Philosophy, Psychiatry and Psychology* 5 (3): 203-207.

Wakefield. 1992. 'The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values' 47 (3): 373-388.

Wakefield, Jerome. 2009. 'Mental Disorder and Moral Responsibility: Disorders of Personhood as Harmful Dysfunctions, With Special Reference to Alcoholism'. *Philosophy, Psychiatry, and Psychology* 16 (1): 91-99.

Wakefield, Jerome C. 2006. 'What Makes a Mental Disorder Mental?' *Philosophy, Psychiatry, and Psychology* 13 (2): 123-131.

Part A

Paper 1 - Ethical decisions in the classification of mental conditions as mental illness

Abstract: Since the early 1990s, the philosophical debate over broad accounts of mental illness has stalled. Whilst there is still unresolved tension between mixed and medical models of mental illness, bioethics appears to be moving from a naturalistic account of mental illness to one in which illness is determined by applying an evaluative notion of function. Nonetheless, existing models often underestimate the role of social norms in defining illness. Most importantly, such models have paid inadequate attention to the relevance of wider philosophical assumptions about the objectivity of ethics and the concept of personhood to our 98understanding of illness. I will attempt to demonstrate that these concepts are integral for differentiating mental illnesses from the vast array of irrational and pre-rational drives and personality traits for which we usually wish to hold the bearer morally responsible. In emphasising the normative component in accounts of mental illness, I am not attacking psychiatric expertise, but rather endeavoring to bring philosophical discussion closer to the actual, informal, decisions that psychiatrists (in particular forensic psychiatrists) regularly make when asked to determine someone's moral responsibility for a mental condition.

It is vital to note that there may be *many* different senses in which 'mental illness' may be defined, ranging from something that justifies adopting the role of a patient with corresponding moral claims to increased rest and medical treatment (at least in those societies with tax-funded healthcare) also known as 'the sick role', to perhaps any unpleasant condition that a psychiatrist is capable of treating. *I am interested only in a particular concept of mental illness: illness as a moral concept that justifies the deprivation of moral responsibility and autonomy*. I explain this in greater detail in the

schedule to this dissertation, ‘Changing functions, moral responsibility and mental illness’ (Edwards 2009). As explained in the introduction, I have chosen to amend my previously published works almost entirely by omission (mostly to minimise repetition, but on some cases for readability) and necessary associated additions, with no conceptual additions to these works. This is so I can honestly state that they are essentially the same works as those I list as publications arising from this dissertation. The closest I come towards insertion is the occasional reference to later papers in the dissertation that will clarify an important detail.

In the context of the broader dissertation, this paper presents the initial step in demonstrating that the justifications for psychiatric paternalism rest on evaluative, rather than objective, concepts of illness and competence. In this paper, I identify the first of several moral choices that define the effective scope of personal autonomy. The full relevance of this paper becomes clear in the last work of Part A, ‘Mental competence and its limitations’, in which I discuss the tension between the aim of social inclusiveness and the requirements we impose upon the recognition of others as equals in the recognition and giving of reasons for action. This paper demonstrates that mental illness is a normative concept, not only in terms of separating harmful from harmless dysfunction, but in determining what mental variations amount to dysfunctions. It is a lesser step in the scope of this dissertation, than the similar finding that mental competence is also an evaluative concept, but one that will be used to show that the orthodox bioethical approach to medical paternalism, i.e. the doctrines of informed consent and mental competence, are flawed from their most basic foundations. It is the first step in demonstrating that the attribution of moral agency is inescapably a social and moral decision (albeit one shaped and informed by medical science), and one that cannot be ‘outsourced’ to the medical classification of illness.

Ethical decisions in the classification of mental conditions as mental illness

1. Beyond taxonomy – the importance of developing a satisfactory philosophical account of mental illness.

Since the time of Plato, philosophy has harbored a somewhat obsessive desire to explain the classifications we apply to things in the world. This has often proved a futile task – many of our classifications contain an inescapable arbitrariness and we are none the poorer for it. By contrast, our search for an accurate account of mental illness is a response to urgent problems of a very practical nature, as well as ethical reflection upon our legal and social rules. Amongst these practical problems are questions such as 'When should we excuse someone from criminal liability on account of his or her mental state?' and 'What kinds of mental condition should bring a person within the scope of the state's involuntary psychiatric treatment programs?' At a more reflective level we need to determine what kind of mental condition will impair a person's responsibility for his or her actions.

Western nations typically contain legislation that implies a special relationship between mental illness and responsibility – i.e. the fact that someone has a mental illness provides at least partial grounds for being exempt from criminal responsibility and modifying their right to refuse treatment. For example, it is common to allow that a person can be detained for psychiatric treatment if she (a) has a mental illness, (b) refuses to consent to treatment, or is incapable of consenting *and* as a consequence of the illness and (c) is in danger of harming their reputation, family relationships or financial security.[i] Where such legislation defines mental illness, the definitions are typically broad and do not differentiate among types of illness (i.e. there is no explicit distinction in the legislation between, say, schizophrenia and depression). Thus the

first question in determining if a person's rights are to be affected by such legislation is that of whether the person has a mental illness. Ordinarily, self-harm to reputation, family relationships or financial security are not considered adequate grounds for paternalistic interference by the state. The classification of mental illness has serious practical repercussions – if a person harms her reputation due to mental illness that person may be subject to involuntary psychiatric treatment, if a person harms his or her reputation for other reasons that person can do so free from interference.

This article is not an inquiry into the term 'mental illness' as it is used in ordinary language. There is no guarantee that such usage is free from arbitrariness, nor that it should be. Nor is it, strictly speaking, an attempt to define 'mental illness' as a term in medical theory independent of its relevance to medical *practice*, although that project is closely related to my purpose. Bioethics is interested in 'mental illness' as a *normative* label – one that *ethically justifies* certain consequences, such as the ill person adopting the role of patient and social denial of that person's responsibility for some of the person's actions. Arbitrariness in this sense of 'mental illness' has serious consequences. A significant proportion of the people who are subject to involuntary psychiatric treatment, or are considered by mental health professionals to be suitable candidates for such treatment, dispute the claim that they have a mental illness. Involuntary psychiatric treatment of a patient who falls into this category can only be ethically justified if we have some non-arbitrary response to the denial that his or her mental condition is truly a mental illness. The adequacy of our response is important to both the treatment of ill and vulnerable people and the protection of individual liberties, and so this is not a matter in which we can comfortably adopt a policy through a loose political and social compromise. It is in this context that I attempt to expand the link between the concept of 'illness' and the broader debates in philosophy

about ethics, meta-ethics and personhood. In particular, an account of mental illness that is capable of supporting the practice of involuntary psychiatric treatment as we know it – i.e. of detaining and treating a patient who rejects the classifications underlying his or her diagnosis – requires us to accept some degree of objective ethical truth. I also draw from Szasz's comments about proper function to put forward an account of mental illness that focuses on dysfunction of personhood. However, this is not an anti-psychiatric account – I hold that mental illness is *real* illness and that it is the right kind of concept to provide a plausible basis for the social practices of involuntary psychiatric treatment and diminished criminal responsibility. There *is* a non-arbitrary justification for our classification of mental illnesses, but it is based in ethical truths rather than value-free features that are unique to mental illnesses.

In Section 2 I will review the recent history of philosophical attempts to define 'mental illness' and where such attempts have led us. In Section 3 I will revisit the work of Thomas Szasz. Whilst I disagree with Szasz's account of mental illness, I will explain how some of his criticisms of institutional psychiatry can be reconstructed in a manner that gives further reason to reject the claim that mental illness involves dysfunction in the biological design imposed by evolution. Rejecting Szasz's alternative explanation that mental illness is merely a failure to comply with social norms, I will propose an account of mental illness wherein a dysfunction of personhood (i.e. a 'personal dysfunction') is necessary for mental illness. This notion of 'person dysfunction' is explored in Section 4. Though value-laden, this is still an account of *real* illness consistent with the recognition by Fulford (2000; 2001), Amundson (2000) and Nordenfelt (2007) of normative components of dysfunction in general illness. In Section 5 I will observe that there are many mental conditions that are both rationality-impairing and harmful, but aren't usually considered to be mental

illnesses. I will argue that the distinction between such conditions and mental illnesses turns upon the relevance of the mental conditions to one's ethical framework. Mental illness is not a label that picks out a set of consistent qualities in a mental condition; it is label that stipulates how people *should* respond to the condition, and in particular whether they should respond morally or medically. Consequently, the qualities necessary for mental illness will vary depending on the ethical beliefs that are relevant to determining the appropriate response to mental conditions with those qualities.

2. Philosophy and mental illness: our current state of affairs

The past 50 years has seen some degree of both convergence and ideological shift in philosophical debate about the nature of the concept 'mental illness'. The 'anti-psychiatry movement' (most prominently Szasz, Laing, Goffman, Bateson, Scheff and Foucault, though as noted by Sedwick (1973, 76), these authors don't represent a singular school of thought) that stirred so much debate in the 1960s and 1970s has largely given way to theories that accept (or at least do not explicitly reject) that mental illness is *real* illness while debating the role of normative judgments in our classification of conditions as illnesses. In this context, 'real' illness has typically been defined by reference to the features of physical illness (Pickering 2003, 244). This debate reflects the acceptance of 'function' as the concept most central to this classification, whilst revealing the lack of consensus about whether that concept can be adequately applied without normative assumptions.

Naturalistic theories, the most prominent being those of Boorse (1977; 1975; 1997) and Wakefield (1992), claim that the dysfunction relevant to illness can be identified as an objective fact independent of values. Boorse defines illness in terms of biological function, being the typical contribution of an organ or process to the organism's

survival and reproduction (Boorse 1977, 542). Wakefield describes the 'natural function' of a bodily organ or process as being 'an effect of the organ or mechanism that enters into an explanation of the existence, structure, or activity of the organ or mechanism' (Wakefield 1992, 382). Nonetheless, the naturalistic claim does not entirely deny the involvement of value-judgments in medical practice. Boorse and Wakefield acknowledge that the appropriate *response* to biological or evolutionary dysfunctions, such as accepting or adopting the 'sick role', excusing anti-social behavior or providing treatment, *are* determined by value judgments, and for Wakefield at least the concept of 'illness' itself includes the application of such value judgments. This is unsurprising – there isn't a strong social consensus on the evaluative implications of mental illness and so it would be, to say the least, challenging to derive the 'ought' required to justify the alteration of people's rights to liberty involved in detention for involuntary psychiatric treatment, from the 'is' of mental illness without some normative component (Fulford 2001, especially at 82).

Contrary to the naturalists are the more thoroughly normative conceptions of illness and its relevant dysfunction.[ii] Under these accounts, normativity is not merely applied *after* the identification of a dysfunction, sorting out benign dysfunctions from illnesses, but is part of the relevant notion of dysfunction itself. Such accounts have proceeded by claiming that the naturalists cannot avoid implications about values in their *use* of terms denoting illness (Fulford 2001), that dysfunction in the manner relevant to health is determined by one's ability to achieve one's important goals (Nordenfelt 2007, 7) or at least strongly influenced by a patient's desires and values (Fulford and Colombo 2004), and that the very concept of 'normal function' (as opposed to function comparative to normatively derived standards) is fundamentally flawed (Amundson 2000).

For sake of brevity I will not critique the entirety of the debate between normative and naturalistic accounts, but my account is clearly a normative one. Thus while the following section provides some reason for rejecting the naturalistic theories, I do not intend it as a stand-alone justification of the normative position. Instead I attempt to offer a line of reasoning that is consistent with much that is said by earlier normative theories, but that provides further guidance regarding how the normative judgments involved in the application of the concept of illness are to be made and how the claims of the normative position are to be integrated into the broader philosophical debates about ethics and personhood.

3. Revisiting Thomas Szasz's argument from function

One doesn't need to accept Szasz's anti-psychiatric claims in order to find merit in his criticisms of the orthodox institutional psychiatric account of mental illness. In the following pages I will attempt to reconstruct an argument that Szasz makes in his response to Chris Megone's attempt to illuminate the concept of mental illness using the concept of the Aristotelian *weltanschauung* (Szasz 1998), but which doesn't appear in Szasz's main published works. Unlike most of Szasz's arguments, this one proceeds from the theoretical concession[iii] that illness can be defined objectively as deterioration in function, and requires little adaptation to be relevant to the notion of dysfunction used by Boorse and Wakefield. In reconstructing Szasz's objection I reject his conclusion, and instead suggest an account of mental illness that involves dysfunctional personhood rather than mere social deviance.

Szasz begins by noting that for physical illnesses there is usually little difficulty in identifying the function of the impaired organ or process. Further, Szasz accepts that it

is possible (at least in principle) to identify the function *objectively* in the context of the human body by reference to the biological design imposed by natural selection. Szasz accepts that mental illness, like physical illness, refers to dysfunction. The difference is in the type of function. He claims that because mental illnesses concern *people* rather than biological organisms, they indicate social dysfunction rather than biological dysfunction.

As discussed in the previous section, there are numerous theories that don't posit biological dysfunction as a necessary component of illness. Nonetheless, dysfunction of some form or another is the current gold standard by which illness is identified. This is no accident – the concept of dysfunction can be comfortably applied to organs, processes and organisms as a whole, and may well be the only concept that can successfully encompass the enormous variety of conditions correctly described as illnesses (Wakefield 1992). Given that the naturalist and normative schools of thought both apply the concept of dysfunction it is worth examining Szasz's claim and its implications.

If, following the work of Fulford, Amundson and Nordenfelt, we believe that the concept of function that is relevant to illness is itself a normative concept, there is good reason to take seriously Szasz's claim that mental illness affects *people* not *organisms*. The common feature of the normative theories of medical dysfunction is that the dysfunction is determined by reference to some purpose that the bearer is evaluatively judged to have an important interest in. In turn, we might say that dysfunction is determined *against* those purposes that are the most authentic goals and values of the patient – her true and most important goals and values, rather than those imposed by illness, weakened will or indoctrination.

Two issues that any such account must face are the multiplicity of contrary purposes an indivudal may adopt, and the possibility that illness itself has altered a patient's values and purposes. In the third paper, 'Mental competence and its limitations' I discuss how the question of how some goals and values are more central, and hence more important to a person than others, whereas in the 5th paper 'Beyond mental competence' (Edwards 2010) I develop an account of personal identity that provides a theoretical basis for the common-place suspicion that a value may be highly central to an individual and yet be an inauthentic imposition of illness. Both challenges can be met, and are met, though they fall outside the matters that can be productively discussed in this topic. We may proceed, then, with the idea of proper purpose being designed by human usage and intent.

For most physical illnesses, this purpose will be identical to the biological function suggested by the naturalistic theories, as there is little room for variation in the interest that a creature is adjudged to hold in most of its bodily processes – a heart attack has the same relevance under all plausible normative systems (Fulford 2001, 82). Survival and reproduction holds a special place in our understanding of the biological world. Survival is a pre-requisite for the fulfillment of all other goals, and evolutionary theory has strongly encouraged us to judge all creatures as having an interest in reproduction. This need not be the case. If, rather than discovering natural selection, humanity adopted a religious worldview in which animals had a special purpose that they were created to fulfill, it is entirely plausible that we would define their health primarily by reference to that purpose, with reproduction taking a lesser importance. But, irrespective of counter-factual possibilities, we *do* judge all creatures to have an important interest in survival and reproduction (albeit for people this is a *defeasible*

interest, as explained in the next paragraph), and so identifying physical illnesses by reference to biological dysfunction is often consistent with a normative account of dysfunction. Hence, it stands to reason that dysfunction in the mental processes required for survival and reproduction are considered illnesses.

Nothing in the discussion thus far implies that mental illness can only involve people, rather than human organisms. If an organism requires various mental processes for survival, it is likely that we would judge the organism to hold an important interest in those processes. However, the interesting thing about *persons*,[iv] and possibly other things that have sophisticated mental lives, is that we value things other than survival and reproduction, and for the most part we evaluatively judge that other people *should* value things other than survival and reproduction (e.g. happiness, or fulfillment). For example, we accept that the interest that gay couples hold in having emotionally and physically satisfying relationships is more important than, and hence overrides, their interest in heterosexual reproduction. Our recognition of mental illnesses reflects this judgment: even though conditions such as depression and agoraphobia can lead to impaired physical health and even death through suicide, such impact upon survival and reproduction is not intrinsically tied to those conditions. Most importantly, our recognition of such conditions as mental illnesses does not rely on their impact upon physical health – we would not dismiss such conditions simply because we were confident that the sufferer in question was in no danger of physical harm. Our 'vital goals' and interests therefore must include the mental and social functions that are alluded to in illnesses such as 'social anxiety disorder' and 'compulsive disorder' – i.e. they must include the social and personal interests that determine a person's 'quality of life'. Goals of this nature cannot be applied to mere organisms: we have no reason to suggest that a mindless organism has an interest in its quality of life, and such a

suggestion may even be nonsensical given that application of concepts involving quality of life and social aims seems to imply some form of mental existence.

If mental illness refers to a dysfunction in a person, this raises a rather timeless question: how does one determine what the function of a person is? For Szasz, of course, such a concept lacks the objectivity he believes is necessary to an adequate account of illness, and can only be cashed out in terms of social, moral or legal concepts with no direct medical application. In the following paragraphs I suggest that in the context of mental illness, the idea of a 'dysfunctional person' should be taken literally. It doesn't refer to a person who fails to meet their social or moral functions, but rather to one who is unable to fully function *as a person* due to impairments in the processes and capacities that are necessary for being a person.

There are three reasons for adopting this narrower meaning of 'dysfunction in a person' as a starting point in an account of mental illness. Firstly, and most trivially, it seems that the conditions we commonly label as mental illnesses can indeed be described as dysfunctions of personhood. Agoraphobic people are dysfunctional because their terror of open spaces is so great that it impairs their ability to choose to go outside (i.e. a loss of agency). Clinically depressed people are dysfunctional because their ability to interpret and emotionally respond to their circumstances in a manner appropriate to their values and belief structures is impaired. Someone experiencing massive personality change or an apparent multiple personality disorder (noting the dominant, and growing, medical skepticism about this diagnosis) is dysfunctional due to a breakdown of continuity or narrative.

Secondly, this understanding of mental illness explains why we differentiate mental

illness from mere deviant behaviour – someone is only mentally ill if they possess the requisite mental symptoms. Thirdly, and most importantly, this is the only account that can explain why mental illness could conceivably be relevant to autonomy and responsibility. It is commonly held that autonomy and responsibility are qualities of personhood – they apply *to, and only* to persons (Feinberg 1986, 270). This does not mean that personhood is a pre-requisite for rights generally (e.g. Marquis 1989). However, rights to autonomy are a special case as they require some actual capacity for moral, psychological and behavioral autonomy in order to be effective. We are responsible for our actions because we have characteristics such as (some sufficient) continuity of character (Schechtman 1996) and capacity for rational agency. Obviously there are biological processes that underlie such capacities, and one could describe those processes as part of the human organism. However, the mere fact that a biological process is impaired is not sufficient for reduced autonomy or responsibility – we don't consider declaring someone mentally incompetent because they have leukemia or cholera. It is the effect that the impaired processes have upon one's personhood - i.e. one's existence not just as a living organism but as a social being with a mental life that is capable of being the subject of rights, duties and responsibilities - that is relevant to autonomy and responsibility.

4. What is 'dysfunctional personhood'?

The function of personhood is not forced upon us by our existence as a particular biological organism. Whilst it is unlikely that anyone could *not* have an interest in the various mental capacities that are often considered to be a necessary part of fully functioning personhood (e.g. capacity for rational agency), that doesn't exclude the possibility of personhood having other functions contingent on our beliefs and values. In this sense, determining the function of one's personhood is more akin to

determining the function of one's body as a whole, rather than of one's organs and bodily processes which can only be changed by altering one's physical existence as an organism (e.g. by replacing an organ with an artificial organ). One's personhood, like one's body as a whole, is something that can be applied to the pursuit of projects and lifestyles and such application may broaden what one demands of one's personhood before it can be considered fully functional.

In a culture where religion provided the dominant understanding of the world, the function of a person could be explained in a very different manner to the explanation given today. For example, if it was commonly held that a person's purpose was to live a good life in accordance with the rules of God, then the functionality of one's personhood may be determined partially by reference to that goal. Any predisposition that made it unusually difficult for the person to avoid breaking those rules would be considered a deviation from proper functioning. If that society was inclined towards seeking medical solutions for such problems, then that dysfunctional predisposition might be considered an illness.

Hence medieval societies might consider predisposition towards deviant lifestyles, such as deliberate childlessness or homosexuality to be dysfunctional. Initially this might be viewed as an inherent moral flaw, but later as parts of society moved towards a medical view of humanity, deliberate childlessness and homosexuality were considered mental illnesses (Szasz 1974, 38). Today such conditions are no longer credible mental illnesses. This isn't merely due to social tolerance - even people who oppose homosexuality and deliberate childlessness usually classify them as ordinary immoral choices rather than mental illnesses. It is due to our liberal conception of personhood, i.e. our narrowing of what one requires to be a properly functioning

person. This is not to say that we all accept a liberal meta-ethics or even a liberal ethics, and there is a need for caution regarding the meaning of 'person' and 'function' in this context. Among Western philosophers, MacIntyre (1981), Taylor (1994) and their supporters have produced powerful criticisms of individualist liberalism, and I must confess a significant degree of sympathy towards Taylor's position. Further, their basic common claim – that liberalism is *wrong* in implying that we can understand what it means to live well as a human being without a developed and objective account of 'the good' – mirrors a worldview that has never disappeared entirely among the general population of any culture. Most people do not accept moral subjectivity or cultural relativity, and many reject the idea that there are even different equally valid ways of living a good life. Such people can still advocate tolerance, whether through a belief that tolerance forms part of the unitary 'good life' that they believe in or through a belief that people should be allowed to 'make their own mistakes' when they aren't harming others. What is important for present purposes is that even these intellectual and cultural authoritarian views still use a liberal notion of personhood. Under such views, living contrary to 'the good life' is not the result of an ailment afflicting a person, but is a bad choice for which one is responsible. There is certainly a sense in which, for MacIntyre, a developed notion of 'the good life' is required in order to determine whether someone is living well as a person. However, the personal and social goals that comprise the good life are not pre-requisites for personhood *in the sense that is relevant to illness and responsibility*. I am differentiating two types of liberal claim here: (a) our existence as human beings and as people does not impose upon us a particular conception of the good life, and (b) deviation from the good life is an immoral or otherwise wrong choice made by a person, rather than a failure to properly function *as* a person. The authoritarians reject (a) but still accept (b). This need not be set in stone. It is in a sense open to the authoritarians to argue that failure

to live certain aspects of the good life is such a frustration of one's important interests that it amounts to a mental illness akin to depression. However, such a claim would have major repercussions upon other aspects of one's ethical system, in particular those determining personal responsibility for one's voluntary actions. This relationship will be explored in the next section.

Whilst we no longer allow our views about what comprises a good life to establish requirements for functional personhood, the requirements for functional personhood haven't been completely eliminated. A person *is* expected to have certain mental capacities, e.g. capacity for rationality and conscious thought. Some philosophers of mind have also argued that personhood also requires certain properties such as sufficient continuity of consciousness (e.g. Tooley 1972, 57; Schechtman 1996, 94). It is accepted, however, that one's personhood isn't rendered dysfunctional merely through deviant behaviour or even a predisposition towards such behaviour. Consequently, in our culture, behaviour alone is insufficient for mental illness – the only way that a person can be dysfunctional (and hence mentally ill) is if he or she suffers impairment to those mental capacities and mental properties necessary for being a person.

5. Differentiating mental illness from rationality-impairing character traits on the basis of ethics

When discussing functional personhood in the context of mental illness, it is common to emphasise the importance of the capacity for rational agency. The proper definition of rationality in this context has been the subject of considerable controversy. Some bioethicists follow objectivist theories of rationality, claiming that some decisions are inherently irrational just because they are contrary to the person's relevant interests

and the applicable moral or social values given the person's interests and knowledge of the surrounding circumstances (e.g. Culver and Gert 1982; Culver and Gert 1990; Gert 1990). Mental illness has also been associated with deprivation of mental *competence*, a concept similar to voluntariness that involves a person's ability to adequately understand a decision, the relevant circumstances and the importance of the potential consequences of one's actions (e.g. Buchanan and Brock 1989; Kleinig 1983, 100–141). Hence I wish to emphasise both 'rationality' *and* 'agency' – i.e. in talking about rationality-impairing traits I am referring to those which (a) tend to cause a person to act contrary to their interests without an adequate reason for doing so, *and* (b) impair a person's ability to decide competently and voluntarily, e.g. by disrupting one's cognitive abilities. I realise that neither of these descriptions are without controversy. Rationality is the subject of such an enormous quantity of philosophical debate, that it is likely impossible to describe without begging some questions, but I do not intend to urge the reader towards a singular comprehensive theory of rationality in this section.

Despite the common emphasis on rational agency, personhood – or at least human personhood – is marked by drives, instincts and dispositions that can at times seem harmful and utterly inconsistent with rationality. Nonetheless, we don't usually consider such dispositions to be either signs of dysfunctional personhood or a species of mental illness. I will argue that such conditions cannot be distinguished from mental illnesses by reference to the objective qualities of the conditions alone. Instead one must consider the normative effect of labeling a mental condition a 'mental illness', specifically that to do so is to declare that people should respond to the condition medically rather than by attributing moral responsibility. Whether such a declaration is appropriate will depend upon what broader ethical framework is presupposed, and so to the extent that different societies can give different acceptable answers to ethical

questions, the relevance of specific conditions or specific qualities of conditions to the label 'mental illness' may vary with different ethical belief systems.

Consider a couple living in a war zone during a bombing raid, frozen in fear under a precarious cover just a few meters from a bomb shelter that would greatly increase their chances of survival. In one sense the couple aren't thinking rationally – they would be better off running to the bomb shelter, and they are aware of this, but they aren't capable of thinking in a sufficiently ordered fashion to take action. However, it seems bizarre to call their fear a 'mental illness'. The problem goes a lot further than basic emotions like fear or anger. There exists a whole array of mental conditions - greed, jealousy, hatred, misery, loneliness, racial prejudice, shyness and nervousness to name a few - that on occasions seem to impair our rationality, are sometimes considered ton egatively impact our wellbeing and that often fall outside of our ability to control as rational agents, yet aren't usually considered mental illnesses. Some of these conditions (e.g. greed) might be portrayable in terms of a Humean account of practical rationality, as a basic desire which is necessary to motivate a person to action. Other conditions, such as nervousness, don't fit that account so readily.

More importantly though, there are numerous examples of common human behaviour that cannot be described as rational under any plausible account of rationality. Consider the example of armed robberies conducted under circumstances where the potential takings couldn't possibly justify the risk to the perpetrator (Mouzos and Borzycki 2003), or people who repeatedly commit assaults under circumstances where any rational observer would know there is no chance of avoiding arrest. These events can be explained by reference to personality traits of the people involved – that they are hot-tempered, shortsighted or lack consideration for themselves or others – but that

only confuses the situation further. If such conditions predispose someone towards intermittent but massive and harmful lapses in rationality, then why don't they count as harmful dysfunctions of personhood in the sense required for mental illness?

The distinction can't be explained by such conditions being statistically less exceptional than mental illnesses – the mere fact that an illness is widespread doesn't stop it from being an illness. Nor can it be explained by claiming that such conditions require an environmental trigger before affecting one's rationality – the same could be said for illnesses such as post-traumatic stress disorder or agoraphobia. Neither can the severity of such conditions differentiate them from mental illnesses – fear can cripple someone's thinking as effectively as depression.

The explanation requires an appreciation of something that has been sadly overlooked in the conceptual analysis of mental illness – that the decision whether to classify a mental condition as a dysfunction is partially an ethical decision. 'Mental illness' is not a term that picks out features that are inherent to the relevant conditions – it is a normative label that prescribes the appropriate way of *responding* to a condition. That isn't to say that the appropriate reaction to a condition is independent of the condition's qualities. However, it is just as dependent on our ethical principles and normative beliefs.

One must also understand the difference between *personhood* and *character*. When we call someone a 'person' in the philosophical sense, and then ask what comprises their personhood, this could mean two different questions. We could be referring to the kinds of qualities necessary to be a functioning person – which, of course, is the sense in which we have been using the term thus far. However, we could *also* be asking

about the series of personality traits, beliefs and values that make up that person's persona or character. On the first interpretation of the question the answer would include concepts such as capacity for rational agency and some sufficient continuity of consciousness. The second interpretation of the question would invite answers along the lines of 'has a quick temper', 'empathises well with others' or 'is shy'. To avoid confusion, I will refer to this second concept as a person's *character*.

Looking back at the examples of rationality-impairing mental conditions that aren't mental illnesses, one can see that these are all part of a person's character. Further, they are all things that one has *moral responsibility* for, even though we are not always capable of immediately controlling them (some reactions to fear, or savage bursts of temper, defy the person's immediate control, and even her long-term control may be a difficult and not entirely successful program of anger management counseling, or even medication despite the condition falling outside the moral excuse that goes along with being a direct symptom of mental illness in its morally significant form. I return to this apparent contradiction in the paper 'Mental Competence and its Limitations' (only a contradiction if we accept that moral competency and responsibility, to be value-neutral qualities that we meet through our mental capabilities). This doesn't just apply to harmful rationality-impairing character traits – one also has moral responsibility for virtuous character traits such as generosity, bravery and being calm-tempered. By comparison, mental illnesses – like all illnesses – are things which we treat ethically as happening *to* a person rather than being a part of that person's character. People don't bear moral responsibility for schizophrenia or depression, nor for being resistant to schizophrenia or depression.

It isn't always easy to see whether a mental condition should be considered a character

trait or a dysfunction. Consider mental conditions like 'being deeply morose, lethargic and feeling hopeless', or 'becomes anxious when interacting with other people'. By recognising these as descriptions of depression and social anxiety respectively, we know to classify them as illnesses and hence as dysfunctions of personhood for which the sufferer isn't ordinarily responsible, rather than part of the sufferer's character. However, the justification for viewing depression and social anxiety in this manner, and not hot-temperedness and shyness, isn't immediately clear. To confuse matters further, psychiatry on occasions seems to classify phenomenologically indistinguishable conditions differently on the basis of the bearer's circumstances – e.g. depression and grief (Bolton 2001, 183). Some have suggested that the classification process itself is far from objective, and that people of different genders would classify conditions differently (Russell 1994, 250–255). One can also observe that the conditions in both categories are not things that someone can directly control. One cannot become no longer hot-tempered or shy simply by choosing to no longer have those character traits any more than one can cure themselves of depression simply by choosing to no longer have the illness. People can still change their character or dysfunction, but only through means such as counseling, psychological treatment or a self-imposed program of sustained critical reflection and gradual change. Changing one's own character through critical reflection is not a simple matter of exercising direct control – unlike, for example, choosing whether to wave one's hands – but is rather a self-designed therapy. In other words, one needs some form of *treatment* in order to change one's character.

The need for, or availability of, treatment doesn't make something an illness any more than plastic surgery makes a crooked nose an illness. Nor does it erase the moral responsibility that one bears for one's character traits – if a person can't change their

aggressive temperament without counseling, that simply means that they have a moral responsibility to undertake such counseling. Thus dysfunctions like Attention Deficit Disorder and depression can't be differentiated from character traits on the basis that they might only be satisfactorily alleviated through medical treatment.

Nor can the categories be differentiated by the conditions' cause. The relative importance of genetic and environmental causes of various mental illnesses is still the subject of research and debate within psychiatry, but it is safe to say that at least some mental illnesses have genetic causes, at least some mental illnesses have environmental causes and many have a combination of both (Pam 1995). What's more, these same competing theories of environmental, genetic and mixed causes also apply to character traits.

Even combining these factors together leaves us without a useful division. There is no set of qualities that are found in all personal dysfunctions but not in any character traits and vice versa. And this is without even taking into consideration the numerous changes in classification that have been applied to various conditions during the past century. Such changes cannot easily be dismissed as a product of increasing knowledge. People might not have known what the biological manifestation of social anxiety disorder was several centuries ago, but surely they were familiar with the mental phenomena. Again, 'mental illness' is a label that brings with it expectations as to how we should respond to a condition and to someone who has that condition. To diagnose someone as mentally ill is to declare that the person is entitled to adopt the sick role and that we should respond as though the person is a passive victim of the condition. Thus the distinguishing features of dysfunction that we should look for are not a universally consistent set of exclusive qualities, but things that provide the

grounds for the normative claim made by applying the label 'mental illness'. Also, as 'mental illness' is a normative label, it is not sufficient to simply identify the conditions to which it is applied – our interest is in determining what conditions the label *should* be applied to. To that end I will first discuss the consequences of applying the label 'mental illness', and then attempt a preliminary sketch of the qualities that give us reason to view particular mental conditions as mental illnesses.

(i) Consequences of applying the label 'mental illness'

Labeling a mental condition a mental illness is to declare that the sufferer should be treated as a passive patient in relation to that condition. The condition is considered not only harmful in this instance, but in need of long-term removal for the sake of the sufferer's welfare interests. This can be disrespectful of autonomy – it is denying any claim the person has to a right to be flawed, i.e. that the mere fact that she has character flaws is not to be used to challenge her status as a functioning person and the liberty-regarding rights that are derived from personhood. People might also have differing views about whether a mental condition *is* a flaw, e.g. some might believe that honor requires them to be hot-tempered, and to call such a condition a mental illness is to deny the validity of their perspective.

If a mental condition is a mental illness, the sufferers are not (barring exceptional cases of deliberate infection or exacerbation of the illness) entirely morally responsible for having the illness. This divestment of responsibility has three aspects. *Typically*, the sufferer is not held morally responsible for having the condition. The relationship between illness and responsibility here is rather loose - many would argue that there is a certain responsibility held by the heavy drinker who develops gout or the marijuana smoker who develops schizophrenia. Nonetheless, even in these cases there is a degree

of misfortune attributed to the sufferer that is inconsistent with full moral responsibility. The separation of illness and responsibility becomes more pronounced with regards to the symptoms of the illness, including sufficiently causally connected behaviour. Barring the exceptional circumstance where the sufferer is attributed near-full responsibility for the occurrence of the illness, the sufferer is treated as though she or he is morally passive in this regard. Lastly, the symptoms of some mental illnesses are such that the sufferer is partially or fully divested of any moral duty to seek treatment for the illness. This is by no means a necessary feature of mental illness, but as I will explain later in this section the sufferer's responsibility for seeking treatment can influence the classification of a mental condition as a mental illness.

This divestment of responsibility can be interpreted as a limitation on autonomy. If people are socially indoctrinated to view themselves as passive subjects upon whom mental conditions are inflicted, those people are denied the freedom to take an active role in forming their persona. They are denied the responsibility to prevent the condition from forming and the freedom to include it as a legitimate part of their persona. The application of this limit on autonomy seems uncontroversial in the case of acute schizophrenia, but it demands caution in its application to less severe conditions. It has been argued that to treat criminal offenders as ill because of their voluntary actions is to disrespect them in a way that is *in itself* an evil irrespective of its consequences. I refer here not so much to the Enlightenment theories of the 'right to punishment' strongly tethered to the ethical theories of Kant and Hegel, but rather the comparatively modern resurrection of the claim by Morris (1968), who proposes a right to be recognised as culpable and hence responsible (through punishment) rather than treated as ill and lacking in autonomy. Whilst the related contention that there is a 'right to be punished' is far more controversial (e.g. Deigh 1984, 193–201), if a patient

identifies with her mental illness, such that she would choose it if choice were possible, it may be that punishing associated antisocial behaviour is more respectful of her individuality and personhood than declaring that she needs treatment.

Finally we need to consider the important conflicting roles of moral blame and amoral illness in the raising of children inculcated with moral virtues. By labeling a mental condition as an illness, we clarify that it is not a legitimate choice for personal development. If a parent notices signs of mental illness in a child, then she cannot dismiss the condition as being part of their child's personality. However, because people are treated as passive victims of mental illness, there is no means of teaching children responsibility to avoid developing mental illnesses, or to develop their adult persona in a healthy way that minimizes the possibility of illness. By comparison, moral blame and punishment are integral to inculcating values into our children by showing that they must be an active participant in developing their own character. When we punish a child for striking another in anger, or stealing their friend's toy in an uncontrolled impulse, we teach them that they must not only avoid committing rationally calculated evils, but also must actively avoid adopting character traits that lead to evil acts. It can also express a moral rule against actions of that *type* – e.g. punishing violent acts that are committed due to impulsive and irrational anger helps express our broader moral prohibitions on violence.

(ii) Traits that are relevant when classifying a rationality-impairing condition as a mental illness (a preliminary sketch)

To provide a complete statement of the traits that justify applying the label 'mental illness', one would need to consider all of the alleged types of mental illness along with all possible rationality-impairing character flaws as well as utilising and

defending a comprehensive ethical theory with which to consider them. However, I will endeavour to provide a mere preliminary sketch to illustrate the normative nature of the label. The following criteria are not intended to be individually necessary for mental illness, nor comprehensive, but are important factors that warrant consideration:

1. **Is the condition harmful to the person who has it?** Harmfulness amongst rationality-impairing mental conditions is certainly not exclusive to mental illness – conditions such as violent temper can be just as harmful to the bearer as many uncontroversial mental illnesses. However, it would be senseless to call something an illness unless the condition generates a need for treatment – i.e. unless the condition is potentially harmful to the bearer. As such, harmfulness is a necessary but not a sufficient quality of mental illness. Similar positions are held by both Boorse and Wakefield. Both use a purely subjective interpretation of 'harm', which for Wakefield means harmful in the opinion of the patient (Wakefield 1992, 386), whilst for Boorse it means harmful in the opinions of *both* the patient *and* the relevant society. This can only be an adequate account for someone who denies the possibility of *objective* harm, otherwise one is left with the bizarre notion that the accurate use of the label 'illness' is determined by people's beliefs rather than the existence of actual harm. Some terms may well function in such a way, but 'illness' does not – to apply the term properly one needs to discover whether a condition is actually harmful, not simply conduct a survey of peoples' informed beliefs regarding its harmfulness. The measurement of harm will turn upon the broader ethical question of whether people can have objective interests or harms that are independent of people's beliefs about their interests – a question

too complex and too important to be resolved here. Where people disagree about whether a condition is harmful, this question will need to be decided in order to determine whether the condition can be an illness to one person but not the other, or whether instead one person is wrong about whether the condition is an illness. The question is certainly important, but the practical difference between subjective and objective accounts of harm is smaller than it may initially seem – even an objective measure of harm will need to give substantial consideration to the person's subjective interests and values to determine whether that person has been harmed.

2. Is there any reason for legitimising the condition as a character trait that one can choose to develop or maintain? To label a condition as a mental illness is to claim that it can never be a legitimate part of one's personality. Even if a condition is often harmful, there may be circumstances where it is seen as motivating appropriate behaviour for appropriate reasons. The issue here isn't the degree of harmfulness associated with the condition – it is whether there is any way of justifying the harmfulness by referring to other circumstances where the condition is legitimised. For example, some cultures may view a potentially violent temper as enabling someone to express justified outrage at breaches of morality. If such a perspective is ethically permissible, then this would give good reason to view a violent temper as a character trait rather than an illness.

3. Is the condition one which can be discouraged through the inculcation of appropriate moral values during childhood? The role of moral blame in inculcating our children with values gives good reason to teach them to bear moral responsibility for developing a healthy persona, *to the extent that such development can in fact be taught*. There is less reason to

provide children with examples of moral responsibility for avoiding schizophrenia if there is nothing that they can actively do to meaningfully minimise their likelihood of developing the disease.

4. Will applying moral responsibility to the condition help uphold broader moral values in one's ethical system? It may be that even if conditions such as violent tempers, alcoholism or inability to adequately appreciate the consequences of one's actions are caused on occasion by unfortunate genetics, uncontrollable environmental factors or are for some other reason outside the ability of parents to prevent through proper raising of their children (as envisaged in the preceding paragraph), there may be other related reasons for applying moral responsibility to such conditions. Our condemnation of such conditions may be necessary to adequately express our moral values concerning violence, self-control and social responsibility. It may be that even if a violent criminal offender's temper and alcoholism was originally caused by factors outside both his and his parents' control, we must still apply a duty upon him to take an active and morally responsible role in seeking treatment for such conditions, or else we could not truly hold our prohibitions against violence as *moral* rules (as most offenders could lay claim to a rationality-impairing character flaw that motivated their conduct). This would in turn prevent us from teaching our children moral responsibility against such acts.

5. Can one have insight into the condition's effect upon oneself and if so, how difficult is it to take an active role in seeking treatment for oneself? Where a condition itself prevents the sufferer from being aware of the condition's existence – such as in some cases of paranoid schizophrenia where the condition causes a rigidity of thought that prevents the sufferer from being

> capable of accepting suggestions that their fears are delusional – there is no reason to encourage the sufferer to take an active role in altering one's persona, and it would be acutely unfair to expect them to do so. Even in cases where some level of eventual insight may be reasonably expected, there will be less reason to encourage autonomy if the condition impairs the person's ability to even indirectly seek treatment. Conversely, if a person is both aware of the illness and in no way unable to make a reasonable judgment about seeking treatment, failure to treat the condition may make the person morally responsible for the condition's further consequences.

Case example: alcoholism

To illustrate the above criteria (which aren't intended to be a comprehensive list), they can be applied to the condition of alcoholism:

1. **Is the condition harmful to the person who has it?** Yes, there is no plausible basis on which the harmfulness of alcoholism can be denied.
2. **Is there any reason for legitimising the condition as a character trait that one can choose to develop or maintain?** There is no conceivable circumstance where addiction to alcohol could motivate appropriate behaviour for appropriate reasons. Aside from the value intrinsic to autonomy, there is no reason to legitimise alcoholism as a character trait that people should have freedom to choose.
3. **Is the condition one which can be discouraged through the inculcation of appropriate moral values during childhood?** Yes, there is some value in teaching children to actively avoid alcoholism, rather than encouraging a role as passive and blameless potential victims.

4. **Will applying moral responsibility to the condition help uphold broader moral values in one's ethical system?** It may be that applying moral responsibility for alcoholism helps express and uphold broader moral rules about moderation and intoxication. If so, this gives some reason not to classify alcoholism as a mental illness.
5. **Can a person have insight into the condition's effect upon oneself and if so, how difficult is it to take an active role in seeking treatment for oneself?** One would usually expect someone with alcoholism to eventually gain insight into their condition, although it may take some time and some outside encouragement. The condition itself presents a barrier to seeking treatment, although the extent of the barrier is an empirical question and may vary between people. Nonetheless, if there is a type of psychological alcoholism wherein the sufferer is mentally incapable of seeking treatment without outside intervention, that would give very good reason to classify the condition as an illness rather than applying moral blame.

The criteria do not provide a simple calculus whereby one can 'do the math' and thus determine whether something should be labeled a 'mental illness'. Instead the criteria must be considered qualitatively. In the above example, if the sufferer is completely unable to seek treatment without outside assistance, it may be too unfair to apply moral blame for a situation that the sufferer cannot change, and that would give good reason to classify alcoholism as a mental illness. However, if the barrier to seeking treatment is less than a total barrier, then there are strong considerations both in favour of applying the label and in favour of withholding it. One would need to determine which of the interests represented by the criteria were more important – e.g. inculcating values through the expression of moral blame, or protecting sufferers from harm by declaring the condition an illness rather than a legitimate part of their persona.

6. Who decides? Individual and Societal concepts of illness

Insofar as there are objective ethical truths, there will be objective truths concerning the proper classification of mental conditions as mental illnesses. However, where there are multiple ethically acceptable value systems, beliefs about harm and so on, particular conditions may be classified differently by different people or different societies without one view being objectively wrong. If the correct classification of a mental condition depends upon the specific set of ethical values that is presupposed, this raises the question of whether it is the individual speaker's values or a defined cultural set of values that is relevant. There is considerable sense in focusing upon the individual speaker's values – to label a condition a mental illness is to make a declaration about how people should respond to that condition, and so one would expect this to indicate the speaker's assent to that declaration. If the speaker disagreed with the declaration, then we would expect the speaker to say 'most people think that condition is a mental illness, but I disagree' not 'that condition is a mental illness but I think it shouldn't be'.

However, in modern societies the classification of physical and mental conditions as illnesses has major social repercussions requiring a coordinated response. Decisions must be made regarding what conditions will receive state-funded treatment, how to allocate national healthcare resources, what conditions deserve dedicated research funding, what conditions provide a legitimate excuse for taking sick leave and so on. For mental illnesses there are also decisions to be made about what conditions are to be considered as mitigation or excuses for criminal offences and what conditions should be included in involuntary psychiatric treatment programs. Most importantly, societies must determine the scope of involuntary psychiatric treatment and state-

funded psychiatric care. In such contexts we are forced to engage in a societal inquiry into what counts as a mental illness through reference to a system of values that hopefully resembles the population's (mostly) shared beliefs.

The tensions arising from competing values are reflected in the existing normative theories of dysfunction and illness. Fulford (2001, 84) notes that there is considerable *legitimate* variation in people's evaluative judgements about mental conditions, and implies that some degree of democratic compromise needs to be sought in order to satisfy the varying and sometimes conflicting implications that such values have for a person's rights and the provision of treatment (Fulford and Colombo 2004). Amundson (2000), on the other hand, appears to resent and reject the imposition of socially determined concepts of *physical* dysfunction upon individuals, and it seems a natural extension of such reasoning to resist the imposition of the term 'illness' upon people who don't agree that their mental condition is in fact an illness.

Following the above, our current practices of involuntary psychiatric treatment and diminished criminal responsibility turn on some important assumptions about how ethical values should work. Involuntary psychiatric treatment uses a concept of illness that is defined *not* by reference to the patient's subjective goals, nor even by a democratic compromise between patients and the various medical experts involved. Whilst there may be considerable room for patient-doctor negotiation and compromise in many treating relationships, such compromise is not *required* in involuntary treatment programs. Indeed the very purpose of such programs is to act upon the view that there are some interests that are so important to all patients that they must be protected irrespective of whether the patient has any subjective desire to accept such protection. Part of this underlying philosophy is the view that some mental conditions

in some circumstances are illnesses irrespective of whether the patient recognises them as illnesses – e.g. if a person's schizophrenia presents a grave threat to his or her physical or social wellbeing, the schizophrenia is a mental illness even if the person rejects that classification.

As such, involuntary psychiatric treatment requires us to accept at least a minimally objectivist understanding of 'the good', i.e. that at least some normative values are also objectively true. I have claimed that the classification of mental conditions as mental illnesses requires us to make judgments about values and the attribution of moral responsibility. These are *normative ethical* judgments. They may not require us to commit to a particular comprehensive moral system, but they are still decisions that inform and fall within our moral framework of values, responsibilities and interests. We don't need to presume that our ethical judgments are objectively true in order to justify our acting on them. For example, we may instead view ourselves as giving a non-cognitivist *endorsement* of judgments about values and interests (e.g. Gibbard 1990), wherein statements about values are interpreted as personal *expressions* of approval or disapproval. Our personal disapproval of activities such as crime, or flaws such as greed, may by itself give us good reason to take action against such things, and popular disapproval may justify the state legislating against them. However, this seems inadequate as an explanation of why we should impose our ethical judgments and the related classification of conditions as mental illnesses, upon someone who doesn't share those judgments, when our motivation is purely paternalistic. If all we have are two competing claims about whether a condition should be an illness, neither of them objectively right, then this seems inadequate as a justification for us to impose our classification of conditions as mental illnesses and all the ethical judgments that such a classification entails. Instead, when we impose our judgment that a person is mentally

ill, even though the person disagrees that their condition is even *in principle* a mental illness, we are claiming that our judgment is *right* and the person's judgment is *wrong.* That is, we are claiming that the person *is* ill, that the condition *does* threaten certain things that we deem to be interests of that person, and that those things actually *are* ethically valuable and important interests of that person.

This is so even if one supports only a minimalist version of involuntary psychiatric treatment, e.g. one in which the only mental conditions for which a person can be detained and involuntarily treated are those which the person would recognise as illnesses but for the presence of the mental condition itself, or those which the person's 'future self' post-treatment or 'true persona' pre-illness would recognise as illnesses. Our acceptance that the views of the unimpaired persona take priority over those of the person when ill is itself reflective of an assumption about 'the good' – i.e. about the kind of viewpoints that are relevant to determining how 'the good' is defined. Similarly, we do not excuse criminal behavior merely because a person makes a bona fide claim to be mentally ill – the person must be mentally ill in a way relevant to the offence *and in a manner accepted by the community (as represented by legislation) as a mental illness*. To impose or reject the label of 'mental illness' is to impose or reject a person's values, and the former is only justified if we have grounds for the latter.

7. Concluding Remark

My aim has been to explore and elaborate upon the ethical component of the classification of mental conditions as mental illnesses, and in particular the tension between the application of the label 'mental illness' and the preservation of moral responsibility. I have not attempted a comprehensive statement of the ethical issues involved in classification nor the actual correct classification of any individual mental

condition. This is not an attempt to steal expertise away from psychiatry, but rather a more explicit statement of the kinds of questions that psychiatrists consider in their practice but which aren't reflected in philosophical discussion of mental illness nor the laws governing social responses to mental conditions. Nonetheless, moral philosophers should recognise these questions as falling within their core competency, and participate in debate over whether a mental condition warrants a medical or a moral response.

8. References

Amundson, Ron. 2000. 'Against Normal Function'. *Studies in History and Philosophy of Biological and Biomedical Sciences* 31 (1): 33-53.

Bolton, Derek. 2001. 'Problems in the Definition of "Mental Disorder"'. *Philosophical Quarterly* 51 (203): 182-199.

Boorse, Christopher. 1975. 'On the Distiction Between Disease and Illness'. *Philosophy and Public Affairs* 5 (1): 49-68.

———. 1977. 'Health as a Theoretical Concept'. *Philosophy of Science*: 542-573.

———. 1997. 'A Rebuttal on Health'. In *What Is Disease?*, ed. James Humber and Robert Almeder, 1-134. Totowa, New Jersey: Humana Press.

Buchanan, Allen, and Dan Brock. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*. Cambridge: Cambridge University Press.

Culver, Charles, and Bernard Gert. 1982. *Philoosphy in Medicine*. New York: Oxford University Press.

———. 1990. 'The Inadequacy of Incompetence'. *The Millbank Quarterly* 68 (4): 619-643.

Deigh, John. 1984. 'On the Right to Be Punished: Some Doubts'. *Ethics* 94 (2): 191-211.

Edwards, Craig. 2009. 'Changing Functions, Moral Responsibility, and Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 105-107.

Feinberg, Joel. 1986. 'Abortion'. In *Matters of Life and Death: New Introductory Essays in Moral Philosophy*, ed. Tom Regan, 256-293. New York: Random House.

Fulford, William. 2000. 'Teleology Without Tears'. *Philosophy, Psychiatry and Psychology* 17 (1): 77-94.

———. 2001. 'What Is (mental) Disease?: An Open Letter to Christopher Boorse'.

Journal of Medical Ethics 27 (2): 80-85.

Fulford, William, and Anthony Colombo. 2004. 'Six Models of Mental Disorder: A Study Combining Linguistic-Analytic and Empirical Methods'. *Philosophy, Psychiatry, and Psychology* 11 (2): 129-144.

Gert. 1990. 'Rationality, Human Nature and Lists'. *Ethics* 100 (2): 279-300.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings - a Theory of Normative Judgment*. Cambridge: Harvard University Press.

Kleinig, John. 1983. *Paternalism*. Totowa: Rowan and Allenheld.

MacIntyre, Alasdair. 1981. *After Virtue*. London: Duckworth.

Marquis, Don. 1989. 'Why Abortion Is Immoral'. *Journal of Philosophy* 86: 183-202.

Mental Health Act (WA) 1996

Morris, Herbert. 1968. 'Persons and Punishment'. *The Monist* 52: 475-501.

Mouzos, Jenny, and Maria Borzycki. 2003. 'An Exploratory Analysis of Armed Robbery in Australia: Australian Institute of Criminology Technical and Background Paper Series No 7'. Government Research Database. *Australian Instititute of Criminology - Home*. http://www.aic.gov.au/documents/F/6/0/{7BF60AB740-039F-4292-93FB-F5B7FB752671}7Dtbp007.pdf.

Nordenfelt, Lennart. 2007. 'The Concepts of Health and Illness Revisited' 10: 5-10.

Pam, Alvin. 1995. 'Biological Psychiatry: Science or Pseudoscience?' In *Pseudoscience in Biological Psychiatry: Blaming the Body*, ed. Colin Ross and Alvin Pam, 7-84. New York: John Wiley and Sons.

Parfit, Derek. 1971. 'Personal Identity'. *The Philosophical Review* 80: 3-27.

Pickering, Neil. 2003. 'The Likeness Argument and the Reality of Mental Illness'. *Philosophy, Psychiatry, and Psychology* 10 (3): 243-254.

Reznek, Lawrie. 1991. *The Philosophical Defence of Psychiatry*. New York: Routledge.

Russell, Denise. 1994. 'Psychiatric Diagnosis and the Interests of Women'. In *Philosophical Perspectives on Psychiatric Diagnostic Classification*, ed. John Sadler, Osborne Wiggins, and Michael Schwartz, 246-258. London: John Hopkins University Press.

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornwell University Press.

Sedgwick, Peter. 1973. 'Illness - Mental and Otherwise'. *Hastings Centre Studies* 1 (3): 19-40.

Szasz, Thomas. 1974. *The Myth of Mental Illness*. 2nd ed. New York: Harper and Row.

———. 1998. 'Commentary on "Aristotle's Function Argument and the Concept of Mental Illness"'. *Philosophy, Psychiatry and Psychology* 5 (3): 203-207.

———. 2006. 'Defining Disease: The Gold Standard of Disease Versus the Fiat Standard of Diagnosis'. *The Independent Review* 10: 325-336.

Taylor, Charles. 1994. 'Can Liberalism Be Communitarian?' *Critical Review* 8 (2): 257-262.

Tooley, Michael. 1972. 'Abortion and Infanticide'. *Philosophy and Public Affairs* 2: 37-65.

Wakefield. 1992. 'The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values' 47 (3): 373-388.

Endnotes

[i] These statutory criteria are the ones used in Western Australia, pursuant to Section 26 *Mental Health Act* (WA) 1996. The types of harm that are required for involuntary treatment vary between jurisdictions, although those used in Western Australia are certainly not unusual in Western jurisdictions. Most Western jurisdictions have at least some roughly similar legislation allowing the

enforcement of psychiatric treatment (and at times detention) in circumstances which do not give grounds for paternalistic interference in the lives of people who are mentally healthy.

[ii] Prominent examples of the normative account can be found in the works of Fulford, Nordenfelt and Amundson. These accounts typically emphasise the variation in views about illness between people and cultures and seek to explain how it is that people can sometimes *legitimately* hold differing beliefs about whether a condition is an illness. Fulford (2001) notes that whilst the naturalists attempt a value-free definition of 'illness', the use of the term 'illness' is routinely value-laden. He explores the logical transition from the 'is-statement' of 'the patient is mentally ill' to the 'ought-statement' 'we ought to treat the patient'. Fulford argues that in the context of physical illness the discrepancy between value-laden use and value-free definition may be explainable along descriptivist lines. The descriptivist account is that the 'is' to 'ought' transition is legitimate where there is a common consensus about how the subject of the 'is-statement' should be evaluated. For example, one can legitimately reason from the 'is-statement' of 'Max *is* selfish' to the 'ought-statement' of 'Max *ought* to give more consideration to other peoples' needs'. Physical illnesses typically involve the same values irrespective of the sufferer, so there is no problem in deriving the normative responses to illness along descriptivist lines. However, there is no such evaluative consensus regarding mental illness and so the descriptivist argument is inadequate in that context.

The normative accounts' response to such problems is to build evaluative judgments into the definition of illness. However, there is considerable variation between normative theories concerning the nature of the evaluative judgments involved. These differences are not merely procedural, instead having significant implications for the resolution of disputes over the application of the term'illness'. Nordenfelt (2007) appeals to the ability to pursue a person's 'vital goals', which is in effect a normative theory of 'proper' function. If we believe that people's most vital interests are objective interests, i.e. are vital interests independently of whether the person subjectively agrees that they are vital interests, then Nordenfelt's theory suggests that there is a single objectively correct way for the term 'illness' to be applied to a given person, but also indicates that 'illness' may have different applications when applied to *different* people (i.e. if different people have different vital goals). By comparison, Amundson (2000) argues (a) that function involves normative judgments and that (b) the common normative judgments classifying people with disabilities as physically dysfunctional are wrong. Amundson doesn't put forward an alternative acceptable normative notion of function, and so he appears to adopt a subjectivist

standard in order to reject the use of the normative notion of function. Alternatively, he may simply be arguing that the current standard is unjust using an implied standard of proper function that is less discriminatory towards the differently abled. Fulford at times emphasises the subjective goals of the person (e.g. 2001, 82), but elsewhere recognises the relevance of interests imposed by other parties (e.g .Fulford and Colombo 2004).

[iii] The theory presented in Szasz (1998) does not reflect Szasz's broader theories of illness. Szasz's comprehensive text The Myth of Mental Illlness (1974) is more reflective of Szasz's settled views (e.g. Szasz 2006). Szasz believes that the only adequate standard for illness is the presence of a physical lesion, and so the only legitimate mental illnesses are those where it is reasonable to suppose that a related physical lesion exists. As such, Szasz has long argued that mental illness is not 'real' illness but is instead a form of social dysfunction. This is not simply a dispute over classification. Szasz denies that mental illness has the relevance to moral responsibility that is usually attributed to it. That is, Szasz believes that mental illness is incapable of justifying involuntary psychiatric treatment and is also incapable of legitimately excusing immoral or criminal behaviour. He has long campaigned for the abolition of both involuntary psychiatric treatment and defences in criminal law based on mental illness. Szasz is sometimes erroneously associated with more recent movements opposing anti-psychotic drugs and the pharmaceutical focus in psychiatry. Szasz does not oppose the use of anti-psychotic drugs and for that matter has had surprisingly little to say about methods of treatment other than his criticism of the use of force and sedation for the purpose of enforcing involuntary treatment.

In light of the above, Szasz's (1998) discussion of illness in terms of dysfunction is intended to reveal inconsistencies involved in the use of dysfunction as a standard for illness, and should not be interpreted as an abandonment of Szasz's physical lesion theory. Szasz's observations on this issue are interesting in their own right. However, the difference between Szasz's theory and the more dominant (naturalistic and normative) accounts is of a far more fundamental nature than the dispute about how the dysfunction standard should be applied. Putting aside the question of whether mental illnesses are 'real' illnesses, Szasz's denial that mental illness can deprive people of autonomy, voluntariness or moral responsibility seems to turn on an unusual interpretation of the phenomena of mental illness and of the symptoms displayed by the mentally ill. As such I have not endeavoured to provide a comprehensive justification of my rejection of Szasz's account. There have been several comprehensive criticisms of Szasz's

account (e.g. Reznek 1991; Wakefield 1992) and I suspect that those who still follow Szasz's account hold vastly different empirical beliefs and interpretation of phenomena to myself.

[iv] I use the common philosophical definition of a person being not merely a human organism but rather something whose mental life characterises that of normal adult human beings (e.g. Tooley 1972, 40–42). Many philosophers believe that certain rights and duties turn upon the possession of personhood, and so the identification of precisely which attributes are necessary and sufficient for personhood has been a long-standing project of academic philosophy. It is at least theoretically possible to have non-human 'persons', but this is unlikely to be a matter of simply possessing the requisite intelligence. There may be some mental attributes necessary for personhood that are uniquely human, or at least would require a mental life of the exact kind experienced by humans. For example, if a species was discovered that had the same intelligence and self-consciousness as humans, but was not a social species and lacked all the mental concepts required to lead a social existence, it is likely that members of that species would not possess the same rights and duties as persons. On the other hand, species that lack our level of intelligence and the extent of our self-consciousness, but who share other important mental characteristics could arguably qualify for some of the rights of personhood. Of course that turns heavily upon how one characterises 'rights', (e.g. contract theories of rights tend to exclude non-humans from participating, although animals can still be the beneficiaries of rights if the human participants view conservation as being part of their own interests) but the creature's lack of full personhood is not in itself a barrier to the rights in question. For example, many mammals seem capable of experiencing depression, and of experiencing mental trauma when removed from their mother at an overly young age, and hence we attribute to those creatures interests in their standard of living and possibly rights protecting those interests.

Full personhood, however, requires a sophisticated self-awareness – indeed it is difficult to see how someone could exercise responsibility or autonomy without some continuing sense of who one is and what one has done and is responsible for doing. Consequently, many philosophers of mind have argued towards some form of continuity test for personhood, e.g. self-awareness of one's existence as a continuing subject of experiences (e.g. Tooley 1972), narrative unity (e.g. Schechtman 1996) and, to the extent that ongoing personhood implies consistent identity, some form of psychological continuity (e.g. Parfit 1971)**.** Similarly, the moral responsibility arising from personhood implies a need for rational

agency; people are not only thinking creatures, we are agents capable of autonomously choosing our actions. Irrespective of whether 'free will' exists at an ultimate level, personhood requires at least a loose practical capacity for free choice even though, in practice, our choices might often be constrained by our circumstances. This is not to say that personhood requires one to have these capacities at all times - mental illness is just one of many conditions that can limit or deprive a person's mental capacities, and we don't usually think that someone who is temporarily, or even permanently, impaired in such a way is therefore not a person. This, however, may simply be a polite convention – we *do* deprive the impaired person of the rights and duties that arise from personhood (but not the many rights that do *not* require personhood). For example, if a person suffers massive cognitive impairment such that the person is unable to dress or feed herself, whether the impairment was present from birth or occurred through injury we do not apply moral responsibility or rights to autonomy to that person, but we do still apply the rights to humane and dignified treatment, as those latter rights can be enjoyed without full personhood.

Paper 2: Reasons, autonomy and paternalism

Abstract: In this paper, I address the liberal approach to state paternalism in a broad sense. While the study of medical paternalism remains topical, in the broader philosophical study of paternalism and autonomy reached its peak in a series of book and authors during the 1980s, in which state paternalism had the dedicated attention of such luminaries as Feinberg, Arneson, Ronald Dwokin, Gerald Dworkin, Kleinig, Frankfurt and VanDeVeer, to list a portion. Sadly, only a handful of these great authors had the opportunity to apply their study of paternalism in detail to the then emerging field of bioethics. A proper analysis of psychiatric paternalism requires some understanding of this earlier, broader, analysis. Through the doctrine of informed consent, the medical context is one in which the liberal claim has a nearly unique dominance over moral debate. By addressing the liberal traditions that drove discussion of paternalism within broader moral philosophy during the 1980s, I hope to provide a moral and ideological grounding for my later analysis of informed consent and psychiatric paternalism.

To this end, I analyse two alternative philosophical traditions that have sought to ground rights to pursue autonomy at the expense of well-being, restricting paternalistic intervention with the way of living that a person endorses, even where paternalism would successfully promote that person's overall well-being. The account that has dominated philosophical discussion was developed during the 1980s by philosophers claiming a Millian approach to autonomy. This neo-Millian account derives rights limiting paternalism from the intrinsic value of autonomy. I argue that these rights can be more securely grounded in a theory of autonomy that is loosely Kantian in nature.

Reasons, autonomy and paternalism

1. Introduction: Neo-Millian autonomy, rational autonomy and the liberal claim against paternalism

There are a number of, now quite uncontroversial, ways in which paternalism can be self-defeating – i.e. causing more harm than it prevents. The paternalist may be ill-placed to judge what is in a person's best interests, paternalism causes unhappiness, and it impinges upon our ability to learn and develop through the application of our mental capacities to free choice (Buchanan & Brock 1989, 30; Regan 1983, 114; Kleinig 1983, 27-29). However, how should we judge paternalism that is *not* self-defeating? Since the late 1970s, liberal philosophers have sought to establish moral limits on paternalism that apply even when that paternalism would successfully promote the person's overall interests (Feinberg 1983; Feinberg 1984, 52-97; VanDeVeer 1986, 124-127; Kleinig 1983, 27-30; Arneson 1980). That is, that an informed and mentally competent person should be allowed to choose what priority to place on the promotion of her own best interests, such that she can choose to risk or to harm those interests without fear of paternalistic intervention.

To this end, the seminal works on paternalism produced during the 1980s developed an account of autonomy inspired by J.S. Mill's *On Liberty* (1977, original 1859), that I will refer to as the neo-Millian account. At the heart of this account is the claim that (a) mentally competent persons have a moral right to *inviolable* personal autonomy, often in the form of personal *sovereignty*, (i.e. a domain within which one has complete moral authority), (b) paternalistic interference with our voluntary choices violates our personal autonomy, and so (c) paternalism is only justifiable by reference to the involuntariness of our choices or defects in our autonomy. However, in recent decades

even liberal philosophers have come to question whether autonomy is capable of the kind of overriding moral worth that the demand for inviolable personal autonomy supposes. Superficially free choices may disguise an internalised oppression, or a learnt submissiveness, which stands at odds with the motivation for valuing autonomy. This creates a dilemma for those citing inviolable autonomy as grounds for opposing paternalistic imposition of values and lifestyles: unless autonomy requires that we hold certain positive liberal values it could not have overriding moral worth independently of our happiness and well-being, *but* if autonomy is defined in such a way, the overriding value of autonomy itself encourages extensive paternalism.

I believe that the liberal project on paternalism can be better pursued by rigorously examining the paternalists' claim that our choices are unreasonable. Rather than asking what business it is of others to interfere, we should ask 'What makes your value structure more reasonable than mine?' In this paper I explain why, without placing any special moral value upon autonomy, paternalism is only reasonable when it can be justified in terms of the person's own deeply held goals. Such an approach is not entirely new (Scoccia 1990; Brock 1988), and it can be loosely traced through Rawls to a Kantian conception of practical reason (Hill 1989, 97-98; Rawls 1971, 248-250). However, most liberal authors on paternalism have neglected Kantian reasoning as a basis for rights against paternalism, due to their perception that, because it is only concerned with protecting our *reasonable* choices, this approach is too narrow to support the liberal goal of being free to choose to one's own detriment. Properly understood this approach (the 'rational autonomy' or 'Kant-Rawls tradition'), provides for a broader claim against paternalism that is often assumed. The presumed restriction of concern only for our 'rational' exercise of autonomy should be viewed in light of the liberal theory of practical reason upon which it is grounded. The basis of the approach

is one of respect for others' capacity for rationality, and an associated skepticism of our standing to judge the rationality of others' choices by reference to our own values. Properly understood, the 'rational autonomy' account provides for an equally broad interpretation of the liberal claim, and one that I argue is more securely grounded than that provided by the neo-Millian account.

2. Autonomy and the value of freedom

Outside of philosophy, autonomy might well be a synonym for freedom, or perhaps indicate the possession of some broad independence from deliberate intervention. For the liberal philosophers who pursue the neo-Millian account, however, it has become a term of art. It refers to a particular kind of relationship between a person and the shape of her life; one in which she is the author of her life path, such that its course reflects her own goals and choices. It resists precise definition, especially as philosophers disagree over exactly what is required to *achieve* autonomy. However, it is clearly a matter of interpersonal freedoms rather than physical ones. Autonomy does not imply that one is free from causal determination (Frankfurt 1988, 11-25; O'Neill 2002, 29) and our autonomy is not flawed by virtue of ordinary physical limitations preventing us from achieving fantastical desires. Feinberg (1984, 31-44) provides one of the more useful attempts at definition, describing it as the possession of a list of closely related qualities: self-possession (being one's own person), distinct self-identity, authenticity (not being the mouthpiece of other persons), self-determination, self-legislation, moral independence, integrity, self-fidelity, self-control, self-reliance, self-responsibility and self-generation.

We can derive moral limits upon paternalism from the value of free choice, without

making any mention of autonomy. We are usually better placed to judge what is in our interests, and the act of choosing contributes to our learning, development, and happiness (Buchanan & Brock 1989, 30; Regan 1983, 114; Kleinig 1983, 27-29). By depriving the subject of these benefits, paternalism itself is harmful, and the would-be paternalist is morally obliged to weigh this harm against the harm that is to be prevented through paternalism. For some philosophers, that is as strong a claim as they wish to make (e.g. Regan 1983). But others have sought a more liberal stance, wherein liberty rights are not reducible to the promotion of our best interests. On this view, mentally competent persons should be free to choose their own values and ways of living, wherein the maximisation of one's best interests is only one possible goal among many.

3. The 'Neo-Millian' account

3.1 Intuitionist foundations

If absolute limits on paternalism are to be justified by an inviolable claim to autonomy, the value of autonomy must not only be overriding, but also *irreducible*. If the value of autonomy could be fully stated in terms of its contribution to our happiness or well-being then we would lack good grounds for opposing paternalism where intervention would promote those primary values. Thus the Neo-Millian account needs to establish that the value of autonomy is both (a) more important than the prevention of any self-harm, and (b) is *irreducible*, or at least not reducible to the pursuit of happiness or well-being (Feinberg 1984, 64-70; Arneson 1980). Yet this part of the autonomy account has resisted demonstration, resting instead upon what seems to be an appeal to intuition. Feinberg (1984, 52) openly concedes that his account doesn't *justify* a moral claim to personal sovereignty, but appeals to a shared sense of moral outrage at

paternalistic interference with one's own actions. VanDeVeer (1986, 62) takes a similar position in describing the right of competent persons to refuse intervention as 'part of the "deep structure" of ordinary moral thought and of any reasonably comprehensive and coherent theory which, at its core, requires a modicum of respect for individual persons'. Kleinig's (1983, 25) reasoning is a bit more tangible, following Mill in claiming that to refuse to recognise a person's ability to choose constitutes a degradation of that person by denying them the exercise of their faculties of judgment, perception, mental activity, and moral preference.

I have no doubt that many people do, in fact, consider their autonomy to be an overriding interest, and would wish to be free from paternalism no matter which of their other interests depend on it. But this view is far from universal. For some people, paternalism seems acceptable simply because in situations where their judgment is flawed they would want their friends, family or the state to intervene. Some would even say that a good friend is, in part, someone who is willing to paternalistically intervene to prevent you from harm. As such, it is not *intuitively* obvious that our interest in the condition of autonomy, where that condition means independence from unwanted imposition, overrides our other interests. Theories of human nature do not have to be intuitively obvious to be plausible or useful. But lack of intuitive consensus is a problem where one is attempting to build rights upon an alleged common demand, or widely held moral framework.

Moreover, even if we accept that autonomy has an overriding and irreducible value, this does not automatically imply limits upon paternalistic interference with our (competent, informed and voluntary) self-regarding actions. Such an inference requires

that we view autonomy as being fundamentally individualistic, requiring independent rather than communal choice. Some philosophers claim that this view understates the centrality of our relational selves, and question whether independent choice is always more important to autonomy than our interpersonal relationships (Meyers 2005; Oshana 2005; Freidman 2005; Anderson & Honneth 2005). I discuss this point in far greater depth in the next paper, 'Mental competence and its limitations'.

3.2 Undue influence and the requirements for autonomy

The 'contented slave', free from impediments upon his desires only because he has none, is a common point of discussion in philosophical works on freedom. However, there is no need for such a creative example – we are already profoundly familiar with stories of people who profess to be freely pursuing a lifestyle of their choosing, yet are subject to such strict rules and unjust social conditions that their 'freedom' seems to be nothing more than indoctrination. A teenage child whose parents reject all western medicine, who is mature and mentally competent but is subject to parental influence and lacks external social contacts, may fall into this category. Sometimes, the same is said of entire sections of a culture, For example, Oshana (2003, 104) gives a detailed account of the lives of of women living within fundamentalist Islamic culture, which she follows by arguing that these women lack autonomy *irrespective* of whether they accept their status willingly. Or at least that is one view – occasionally the 'victims' are so inconvenient as to point out that our own values are also the product of our social environment, and chastise us about our cultural bias and western ethnocentricity.

Such cases challenge the notion that autonomy is simply a matter of following one's own goals, or living one's preferred way of life. The people in question are self-

determining in that they are pursuing goals that they identify with and would not voluntarily choose to change their way of life. But is that really something that bears the kind of superior moral value that the Neo-Millian account requires of autonomy?

My concern is not that one's goals may be causally determined by one's social environment. Instead, the problem is that if autonomy is to have overriding and irreducible value, it may require that we hold the *right* goals. Lives dedicated to submission, or to unquestioning obedience to a set of rules, seem at odds with the notion of autonomy as part of the good life. A crude approach would be to state that if the moral claims that arise from our interest in autonomy take the form of a right to non-intervention in such cases, then they amount to little more than an excuse to turn a blind eye to oppression. On this view, the victims might be happier than they would be if they didn't identify with and endorse the restricting norms, but there is no *irreducible* value to such autonomy, merely its contribution to happiness and well-being. The approach is crude because our allegation of oppression is a difficult one to make stick – not only would the alleged victims deny the label, but others could argue that it amounts to a mere assertion of cultural superiority.

But we can make the same point without any mention of 'oppression' or similarly contestable value judgments. The people in such examples live a way of life that contradicts the intuitive basis upon which the autonomy account is founded, i.e. that deep moral outrage at others exerting control over their self-regarding choices, values and lifestyle. These people are not reacting as though the 'deep structure' of their moral thought has been violated. In fact, as Oshana's (2003) discussion highlights, *some* seem to welcome and endorse their condition. If we shift the consideration away

from examples as extreme and emotive as the infamous Taliban regime to more familiar intrusions upon autonomy – constant CCTV camera coverage in the United Kingdom, or intrusive stop and search practices in the United States, it becomes clear that a great many people do not share any intuitive moral revulsion at the state's intrusion into their personal sovereignty. Regardless of whether this is genuine oppression, it is at odds with the fundamental basis for the claim that personal sovereignty holds some irreducible and overriding moral value.

It is worth noting that Taylor (1985, 187-210) makes a similar point from a different direction. His concern is that morally meaningful autonomy requires both (a) the development of the mental capacities relating to choice and reasoning, and (b) the availability of credible options that one can choose among. He argues that both of these things require the right kind of cultural institutions and hence if we value autonomy we ought to promote those moral values that are conducive to those cultural institutions. Obviously autonomy requires the right mental capacities, but it is questionable whether this rules out any *existing* social structures – we don't usually deny mental competence to people based on their culture. Similarly, it is difficult to specify what breadth and type of positive freedoms one needs for a choice to be meaningful. Nonetheless, the thrust of Taylor's argument has proven troubling for the Neo-Millian account.

If either concern is true, it is self-defeating to cite the value of autonomy as grounds against the imposition of the values required for autonomy. The enormity of the consequence this has for protection from paternalism is easy to miss, until one realises that these values are those of western liberalism, i.e. one of the most dominant cultural

value systems in the world. In countries such as the U.S. and Australia, this is tantamount to saying that the dominant culture has free reign to paternalistically impose its core values upon members of opposing subcultures. And why stop there? Paternalism is only *one* threat to autonomy, so why not combat the others? Why not work to eradicate cultural and religious minorities (or even majorities) that discourage self-analysis, or that preach the subjugation of the individual to some collective or rule-based authority?

In this manner, the autonomy account slips from being one of liberalism regarding paternalism, to one of broad political liberalism. This may be a non-issue for some, who already see restrictions on paternalism as just one part of a larger liberal project. But for others there is good reason to separate these two forms of liberalism. Concern for unjust paternalism is not limited to supporters of political liberalism. For the many who, like myself, have doubts about political liberalism, the link between the two would be enough to encourage us to seek an alternative ground for limiting paternalism. Rather than an inviolable claim to autonomy giving us the right to enjoy different ways of life, freedom and personal sovereignty are elevated to goals in themselves, at the expense of those lifestyles that do not prioritise self-sovereignty.

4. Principles and paternalism

4.1 Paternalism and harm to self

Paternalism is neither aimed at, nor justified by, prevention of harm, despite the frequent careless statements by some philosophers to the contrary. We don't live our lives in slavish deference to our best interests, nor does anyone expect us to. There are

cases where a person will undoubtedly be *much* better off if she successful,y robs a bank, or escapes rather than turning herself in for a serious crime, or keeps some lost property rather than returning it, and yet no-one seriously suggests that failure to prioritise one's own interests ahead of other goals (such as morality) in such cases warrants paternalistic intervention. Morality is not the only thing that we regularly prioritise above our well-being - we commonly prioritise family, friendship, or even art, sport or science. Whether we gain some evolutionary or personal benefit in a round-about manner from our interest in being moral or goal-driven beings is beside the point. Caring for others might confer some benefit upon us, but if it doesn't then that hardly weakens the reasonableness of our making sacrifices for our children's well-being, or our donating to charities. We prioritise such goals because we believe that they are more important than our self-interest. Such justifications are unnecessary because our self-sacrifice is entirely reasonable. Self-harm is not the target of paternalism, *unreasonable* self-harm is. Thus we can, without any mention of autonomy, derive moral limits on paternalism by establishing what kinds of self-harm are reasonable.

4.2 Practical reason and well-being

Unfortunately, there is no consensus about what choices are reasonable. Some define reasonableness by reference to an allegedly objective set of principles (Gert 1990; Culver & Gert 1982). Given the extent to which people and cultures disagree over what is reasonable, the authority and precise nature of the principles of reason asserted by such accounts are notoriously difficult to establish. Nevertheless, many people seem to find this understanding of practical reason highly intuitively appealing, and there have been some impressive demonstrations as to why objective principles of

reason are not implausible (Parfit 1984, 117-135).

However, paternalism is concerned only with a particular subset of practical reason, being our self-oriented reasons for action. The moral permissibility of paternalism is distinct from the matter of whether there is some reason for intervention that is external to the person being assisted, such as the well-being of a third party, or the upholding of an ideological principle. Even from an objectivist perspective, there remains the question of how our self-interested or self-oriented reasons for acting relate to our subjective desires. In this context, one philosophical tradition is to argue that our self-oriented reasons for action are determined by our subjective network of values and goals (Brock 1988; Scoccia 1990; Rawls 1971, 248-250). This is not the same as cultural or individual relativism – people and cultures can be mistaken about what is reasonable, either by lacking insight into their own deeply held goals, or by erring in their judgment about what things are consistent with those goals. However, these accounts require that self-oriented unreasonableness can only be established by reference to a person's own attribution of rational authority, as reflected in that person's goals and commitments. Thus there is little difficulty establishing the source of the principles' authority – it is asserted by the person herself.

In the following sections, I argue that Kantian reasoning supports this account both by recognising that our goals and values provide us with reasons for acting, and in providing for a skepticism about our standing to judge that others should hold goals and values, or different prioritisations within their network of goals and values, that are currently alien to them. Paternalism involves not merely a theoretical moral stance, but rather the application of a moral judgment through practical conduct. As such the mere

possibility, or even the coincidental actuality, of someone's conduct being unreasonable is not enough to justify paternalistic interference. It is one thing for there to be rules of practical reason that are independent of the internal mental states of any individual. It is another thing entirely for us to have good grounds to believe that our sense of reasonableness is a closer approximation to those rules than the sense of reasonableness held by another person. To have standing to engage in paternalism, our grounds for thinking that another person's conduct is irrational need to be sufficient to overcome the presumption of capacity for rationality that underlies our moral practices.

Consequently, we should not be concerned with what makes something reasonable, but instead with what makes one person's view of what is reasonable superior to that of another person, i.e. what gives *standing* to intervene. Regardless of whether reasonableness is objective or subjective, one can only have *standing* to intervene on the basis that another's choice is unreasonable where the alleged unreasonableness can be derived from the subject's own deeply held goals. To this end I make two related claims. Firstly, that to have the capacity for rational personhood is to be capable of *accessing* the principles that determine what is reasonable. Secondly, that although we may differ in the impediments to our use of our capacity for autonomy – e.g. through different intellect, patience, will-power, and level-headedness – by virtue of our possessing the capacity for rational personhood, we are equally entitled to respect in the sense that claims to rational authority over us must be derivable from shared principles of reason.

4.3 Reasons and capacity

Even on an objectivist perspective, reasons for action cannot be completely independent of people. Principles of reason do not exist 'out there' in the universe. In fact, the opposite is true: principles of reason are an indirect feature of the way that our brains enable us to interact with our environment. We interpret the various environmental data into a form that makes it intelligible, where that intelligibility is itself defined in relation to our needs. Some reasons might be universal, in the sense that any properly functioning mature human should interpret certain inputs as having the same relevance to one's actions. But they are not independent of *us*, the participants in the reasoning process.

O'Neill (2002, 91) says that 'reasons are the sorts of things that we give and receive, exchange and refuse.' What this reflects is that reasoning is a shared enterprise. The capacity for rational personhood requires not only that we can interpret the world and justify our interactions with it so that they are intelligible to us, but that the manner in which we do so is intelligible to *others*. Kant (1998, original 1785, pt.4) gave us the distinction between causal and practical standpoints, wherein we embrace the role of causally determined object and responsible agent respectively. From the causal standpoint I must acknowledge that my choices are subject to external causation and are affected by my limited intelligence and rationality. From my practical standpoint, I give reasons as a rational agent, one that *chooses* and is morally responsible for my choices. To be capable of rationality, our practical standpoint must be compatible with the practical standpoints of other rational people.

Given the nature of rationality as a shared project involving mutual intelligibility, we can only give other rational people reasons for action if those reasons are intelligible to

them (Hill 1989, 97-98). We don't need to agree with their beliefs and values, but if the basic principles from which we derive reasons for action are not compatible with each other, we cannot achieve mutual intelligibility, and so either they or we or both lack the capacity for rationality. Thus, we don't give someone a reason for action by citing a reference to a religion or set of assumptions not shared by that person. We either need to work from common grounds, or our appeal to authority must itself be supported by reasons that are intelligible to our (mentally competent) audience.

4.4 'Sufficient' capacity for rationality, and 'accessible' reasons

The capacity for rationality is a terrible guarantee of good judgment. Having the capacity for rationality is one issue, but actually *exercising* that capacity is another thing altogether. What's more, some people (and all of us, in our worse moments) are subject to characteristics that stand in the way of the proper exercise of our capacity for reasoning, such as stubbornness, short-sightedness and narrow-mindedness. So to say that a person is 'capable of recognising reasons', or can 'access' reasons is not to say that the person will *actually* derive those reasons given sufficient time and knowledge.

Nonetheless, to require that claims about reasonableness be accessible to other persons is to make an important point about the basis for claims about what is reasonable. No person can *unilaterally* determine that another (mentally competent) person's choice is unreasonable. Appeals to some unshared authority, or to the conviction with which we hold our views, cannot be put as reasons in any mutually compatible set of practical standpoints, and are of no use in showing that something is unreasonable. This is not a

denial that some of us are more intelligent or more rational than others, but rather an acknowledgement that being more intelligent or more rational doesn't by itself give our views greater authority. To justify a claim about reasons, we still need to be able to demonstrate the claim in terms of the basic and shared principles that grant intelligibility.

4.5 Reasons and Paternalism

So, what can make a choice to risk self-harm unreasonable? One possibility might be that it will, indirectly, affect others in an unreasonable way. But that isn't our concern for the present topic of paternalism – intervention to prevent unreasonable choices of that nature would be intervention for the sake of *other people*. Another possibility is that some principle, the authority and origin of which is independent of the person in question, declares that we should prioritise our well-being in those circumstances. This kind of claim can be deceiving – it may look like paternalism, but it isn't. The intervention is aimed ultimately at protecting a principle, rather than a person's interests. If, according to some law of morality, it is wrong to risk one's life in order to enjoy drug use or cliff-diving, to enforce this law would be to protect *morality* rather than the person herself. Preventing the harm in question would just be the means by which morality is promoted on that occasion.

There is a problem, then, with accounts of paternalism that define reasonableness as some set of principles that exist independently of any person (e.g. Gert 1990). Aside from sitting uncomfortably with the aforementioned description of what reasons are, such accounts locate the rationale for intervention outside the subject of that intervention. Principles of that kind *cannot* justify paternalism, as they have nothing to

do with person in question. Such accounts do not describe paternalism, but rather the subjugation of an individual to an independent set of rules. That might indeed be justified in some cases, but we ought to be truthful about our motivations for such intervention. I suspect that there is at least a sizeable portion of people for whom 'preventing self-harm to promote morality' or 'to impose common sense' is a less satisfying justification for intervention than 'preventing self-harm for the person's own sake'.

For our intervention to be truly for the person's own sake, and not just preventing their self-harm as a means to some external goal, the unreasonableness needs to come from the person herself. The obvious source is the person's own goals. These reflect the person's own attribution of reasons and authority. But *which* goals are we concerned with? Should paternalism be aimed at protecting a person's *operative desires*, by preventing instrumental irrationality arising from muddled factual judgments or poor information? This would exclude psychiatric paternalism from ever involving mentally competent patients – given the availability of corrective information and time for deliberation, there is no reason why a competent person's decision about treatment would undermine his operative desires.

But our own experience of weakness of will, and similar phenomena, suggests that our operative desires are not always the most important of our goals, even subjectively speaking. The bulk of philosophical writing on this topic has tended towards acceptance that *some* goals and commitments are indeed more central to our self-identity and take priority as reasons for action (Frankfurt 1999, 129-141; Frankfurt 1988, 11-25, 159-176; Valerius 2006; Dworkin 1989; Young 1989; Wolf 1989;

Meyers 2005; Oshana 2005; Christman 2005; Waldron 2005; Bratman 2002), though there is little consensus over how these goals should be determined. The common thread is that one's 'self' is not some essence that pre-exists one's goals, but is instead delineated by one's identification with certain deeply held goals. As such, we attribute more authority to those goals that we identify with deeply. We determine what is reasonable by a choice's consistency with those goals.

As reasonableness here is concerned with goals rather than freedom, there may be values that we are rationally required to hold even if we don't *subjectively* care about them at all. This is because some values may be pre-requisites for the effective pursuit of our deep goals. The most notorious of these is health. Poor health prevents us from effectively pursuing many of our goals, whether we care about our health or not. Moreover, life-threatening conditions endanger almost all of our goals. We cannot effectively pursue the goal of autonomy if we are to be bed-bound and soon dead. The only goals that do not require our survival are those aimed at people or things external to ourself, and those which concern the manner in which we die. It is for this reason that we don't consider it unreasonable to risk one's life to save others, and it is also why religious objections to medical treatment hold such an unusual status. When a Jehovah's Witness prefers death to a blood transfusion, he may well be undermining many deeply held goals, such as his commitments to his career and happiness. However, if he genuinely holds his religious commitments to be the most important thing to him, then this governing goal requires not that he survive, but rather that he uphold the external rules of his religion and that he die in the right kind of way.

Similarly, mental health can be a pre-requisite for other goals. This is most critical

when a person who is currently mentally competent, refuses treatment for a condition that will either render her incompetent or erode her personhood. Psychiatrists practicing in jurisdictions where mental incompetence is a necessary criterion for involuntary treatment, have long complained that patients are locked into a cycle of deterioration, hospitalisation, recovery to the point of mental competence, and then - free to refuse treatment - deterioration once more. There has been some philosophical sympathy towards this complaint (Spellecy 2003; Quante 1999). There nothing gained by refraining from intervention due to the person's goal of autonomy, when it can be predicted with certainty that without imposing treatment the person will be drastically incapable of pursuing that goal,.

We may be able to derive reasons from a person through means other than examining her deeply held goals. The relevance of those goals was simply that they reflect the person's own attribution of authority and reasonableness, and so the person cannor then dispute their authenticity as grounds for determining what is reasonable. It may be, however, that there are other facets of human nature that determine what is reasonable in a self-regarding sense. I am not referring here to such naturalistic claims, as 'good health is required for happiness' – even if true, such a claim does nothing to show that happiness should guide our decisions. Instead, I am referring to normative theories of human nature, that establish what we *should* care about with regards to ourselves. In other words, I am not ruling out the possibility that there are some goals that are simply unreasonable, and others that any reasonable person must hold. It is here that my earlier statements about *standing* to engage in paternalism are relevant. Such 'objective' rules of reason may exist as a consequence of universal facets of human nature, but without some argument from shared principles of reason to that end,

the paternalist lacks good grounds for citing such rules as a basis for intervention. That is, I might be right when I claim that any reasonable person must value art of some sort – but I lack standing to paternalistically impose this view upon another mentally competent person whose goals don't include or require such a value, unless I can justify it in terms of shared principles.

5. Practical reason and undue influence

As noted in section 3.2, the neo-Millian account invites the question of whether we should extend rights against paternalism to those who are living under circumstances of oppression or subject to enormous imbalance in power or information, where they are either unaware or unwilling to recognise their condition. Extending such rights on the basis of the intrinsic value of autonomy is problematic: either such individuals are not really exercising autonomy, or autonomy is less valuable than the liberal claim suggests. Consequently, paternalism on such grounds is tempting, both in terms of societal or cultural oppression such as membership of closed cult-like communities, or in terms of 'saving people from themselves' where they are subject to imbalanced transactions of power or information such as commercial organ markets.

Nonetheless, there is something I find deeply troubling about the notion that those who lack autonomy due to oppression should therefore be legitimate targets of paternalism. It implies that a person's history of being denied autonomy should provide grounds for *further* denial of autonomy. This reasoning exposes the person to denial of autonomy by the same institutions that denied her autonomy in the past. Presumably those advocating paternalism would say that their intervention will be different, and that the

person must be 'made ready' for autonomy once more. The history of such claims, however, is an ugly one: some of the worst cases of systematic racial discrimination in professed liberal democracies are those that have claimed justification on the grounds that paternalistic instruction is necessary to provide the education, information and experience required for 'true' freedom.

Of course, the bona fides of claimed paternalistic motivations for policies such as South Africa's Apartheid is questionable, with fault lying in the imposition of paternalism by race. If so, however, the same charge can be levied against the imposition of paternalism by religion or culture. Moreover, the denial of rights against paternalism to those living with oppression is problematic even when applied on an individualised basis. The claim that a person's endorsement of her circumstances is morally invalid through undue influence, despite an absence of mental impairment arising from such indoctrination, is too often unfalsifiable. Conversely, mental incapacity, possibly including *some* forms of indoctrination as I address in 'Beyond Mental Competence' (Edwards 2010), can be measured and verified. In the absence of such impairment, a person has no means of disproving or verifying the allegation that she is unduly influenced to the point that she lacks meaningful autonomy.

There is no reason in principle why people can't be disempowered by the same cultures and institutions that they identify with and willingly participate in, and Oshana (2003, 104) is right to question the value that autonomy, understood in terms of non-intervention, holds in such circumstances. However, it does not follow from this that paternalism is an appropriate response. The rational autonomy account offers one expression of why we can be *both* concerned that a person is subject to undue

influence by oppressive institutions *and* skeptical of the case for paternalistic intervention. The account appeals to the limitations of our moral knowledge and what it means to take seriously the idea that other beings are capable of engaging in reasoning and of moral responsibility. The Neo-Millian account struggles to address such scenarios because it requires that we accept that a person living the way of life that she endorses really *is* living a life of value. The rational autonomy account allows us to condemn the circumstances of her oppression due to our deep felt contrary moral convictions, while recognising that the strength with which we hold our convictions is not enough to give usn standing to paternalistically intervene in the self-oriented choices of a mentally competent individual. Our moral knowledge is limited and fallible, and when our desires to paternalistically intervene rests upon a clash of convictions, that cannot be resolved by reference to shared authorities or principles of reason, to impose our convictions requires us to either deny the other's capacity for personhood, or to abandon our own claim that our intervention is grounded in rationality rather than a morally arbitrary assertion of authority.

6. The appeal to Personal Integrity

Unedr the liberal accounts of both autonomy and of practical reason, the breadth of protection from paternalism is governed by the robustness of our account of self. Our identification with our motivations for action defines the scope of our self-authorship and authenticity, such that we self-author *by* choosing in accordance with our deep goals and values. It is through our identification with our connative states that we attribute value to them, thereby acknowledging them as reasons for acting. We may view both accounts as alternative (but not mutually exclusive) bases for a common appeal to *personal integrity*: a demand that we be free to live in a manner that is

authentic with regard to our experience of self and our attribution of value.

A shallow self, based solely in immediate experience, identifies only with those goals and values that form one's operative desires. Autonomy and reasonableness would be substantially equivalent to voluntariness. A more robust self allows for the holding of goals and values that are not represented by, or may even be contrary to, our operative desires. With a temporally persistent self, autonomy and reasonableness require that we pursue something more akin to a 'way of life', whereby our more central goals and values are those which characterise that extended way of life. We could choose voluntarily and yet fail to choose with personal integrity if our momentary desires undermine our aspirational commitments.

The rational autonomy account encourages this longer perspective, as most of us already attribute greater worth to our our long-term commitments and persistent values than we do our momentary desires. Nonetheless, the neo-Millian account is unlikely to provide a broader protection against paternalism. Of those who explicitly reject the relevance of examining practical reason when establishing liberal limits on paternalism, their justification is usually that they want to be free to make *unreasonable* decisions – that people should have some authority over their own lives that is not fettered by adherence to imposed principles of reason (Arneson 1980; Kleinig 1983, 26; Feinberg 1984, 96). They wish to ground such a freedom in the claim to inviolable personal autonomy, wherein the scope of that autonomy includes our unreasonable – but nonetheless, competent, informed and voluntary – choices. Autonomy cannot have so wide a scope, and still be something that is capable of bearing overriding moral value. The appearance to the contrary arises from conflating

the contribution of free choice to well-being with the (claimed) role of autonomy as a fundamental feature of the human good.

For autonomy to have such intrinsic value, that value must be located in an autonomous *life* rather than an autonomous *choice* or series of choices. When weighing up the value of alternative courses of action, it seems bizarre to say that the autonomous choice is always intrinsically and irreducibly the more valuable one. Choices don't seem to be the kind of things that can *have* irreducible value – they can be valuable because they assist our learning and development, or because we like the outcome, but not merely by virtue of their existence. For autonomy to be something that has irreducible and overriding value, it has be part of a way of life. When liberal philosophers actually describe autonomy that is exactly how they define it - as something that involves the shaping and design of one's life, the self-determination of one's outcomes and living with integrity and authenticity to one's values (Feinberg 1984, 52, 65-70; VanDeVeer 1986, 62; Kleinig 1983, 25). Yet unreasonable choices *undermine* such a way of life. As Kleinig puts it, our 'lives do not always display the cohesion and maturity of purpose that exemplifies the liberal ideal of individuality, but instead manifest a carelessness, unreflectiveness, short-sightedness, or foolishness that not only does us no credit but also represents a departure from some of our own more permanent and central commitments and dispositions' (Kleinig 1983, 67). For autonomy to include *unreasonable* choices, we would need to dilute the concept away from that of shaping one's life, to include any competent, informed and voluntary choosing. This changes autonomy from a way of life to a mere activity; autonomy is shifted from being (allegedly) an integral part of the good life to something that is valuable only through its contribution to happiness and well-being. In trying to

establish an inviolable right to make unreasonable choices, one would undermine the fundamental basis of the Neo-Millian account.

We should also be careful not to understate the protections provided by the reasonableness account. Autonomy and freedom are capable of being goals just like any other aim. And in the case of those such as Feinberg and Mill, it seems fair to say that they are deeply held goals. Moreover, almost all of us place freedom of choice somewhere in the structure of our deep goals, or else we would not be so concerned about paternalism. This provides a secondary limitation upon paternalism – an unfettered choice might undermine some of our deep goals, but still serve towards one's goal of autonomy. Unlike the autonomy account, however, this does not profess that all of us should place such importance upon autonomy. It permits absolute freedom from paternalism to those unlikely people for whom such freedom is their deepest aim in life, whilst placing no limit upon those whose deep goal structure favours the successful pursuit of other goals (such as health, or family) ahead of their independence.

7. Concluding remarks

Despite practical reason's place at the head of the philosophical table, it remains neglected as a basis for rights against paternalism. Liberal and non-liberal accounts continue instead to be transfixed with autonomy, either as an inviolable moral right, or as something that ought to be balanced against the severity of the harm being prevented. In outlining how substantial limitations on paternalism can be derived from practical reason, I recognise that I have left a great many loose ends unexplored or

undeveloped. My aim in this paper is one of rehabilitation rather than rectification, endeavouring to show that the main concerns that trouble liberal accounts of autonomy may be avoided by instead grounding the liberal view of paternalism in an account of practical reason.

References

Anderson, Joel, and Axel Honneth, 2005. 'Autonomy, Vulnerability, Regonition and Justice.' In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, Cambridge: Cambridge University Press.

Arneson, Richard, 1980. 'Mill versus Paternalism' *Ethics*, 90 (4): 470-489.

Bratman, Michael 2002. 'Hierarchy, Circularity, and Double Reduction' *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. Sarah Buss and Lee Overton, Cambridge: MIT Press: 65-85.

Brock, Dan, 1988. 'Paternalism and Autonomy' *Ethics*, 98 (3): 550-565.

Buchanan, Allen, and Dan Brock, 1989. *Deciding for others: the ethics of surrogate decision making*, Cambridge: Cambridge University Press.

Christman, John, 2005. 'Autonomy, Self-Knowledge and Liberal Legitimacy' In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Anderson, Cambridge: Cambridge University Press: 330-358.

Culver, Charles, and Bernard Gert, 1982. *Philosophy in Medicine*, New York: Oxford University Press.

Dworkin, Gerald, 1989. 'The Concept of Autonomy' In *The Inner Citadel: Essays on Individual Autonomy* ed. John Christman, New York: Oxford University Press: 54-62.

Feinberg, Joel, 1984. *Harm to Self*, New York: Oxford University Press.

Feinberg, Joel, 1983. 'Legal Paternalism' In *Paternalism*, ed. Rolf Sartorius, Minneanapolis: University of Minnesota Press: 3-18.

Frankfurt, Harry, 1999. *Necessity, Volition, and Love*, Cambridge: Cambridge University Press.

———.1988. *The importance of what we care about: Philosophical Essays*, Cambridge: Cambridge University Press.

Freidman, Marilyn, 2005. 'Autonomy and Male Dominance' In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, Cambridge: Cambridge University Press: 150-173.

Gert, Bernard, 1990. 'Rationality, Human Nature and Lists' *Ethics*, 100 (2): 279-300.

Hill, Thomas, 1989. 'The Kantian Conception of Autonomy' In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, New York: Oxford University Press: 91-105.

Kant, Immanuel, 1998. *Groundwork of the Metaphysics of Morals* Gregor, ed., New York: Cambridge University Press.

Kleinig, John, 1983. *Paternalism*, Totowa: Rowan and Allenheld.

Meyers, Diana, 2005. Five Faces of Selfhood. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, Cambridge: Cambridge University Press: 27-55.

Mill, John Stuart, 1977. *On Liberty.* Robson, ed., Toronto: University of Toronto Press. Available at: http://oll.libertyfund.org/title/233.

O'Neill, Onora, 2002. *Autonomy and Trust in Bioethics*, Cambridge: Cambridge University Press.

Oshana, Marina, 2005. 'Autonomy and Self-Identity', In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, Cambridge: Cambridge University Press: 77-97.

———. 2003. 'How much should we value autonomy?' *Social Philosophy and Policy*, 20 (2): 99-126.

Parfit, Derek, 1984. *Reasons and Persons*, New York: Oxford University Press.

Quante, Michael, 1999. 'Precedent Autonomy and Personal Identity', *Kennedy Institute of Ethics Journal*, 9 (4): 365-381.

Rawls, John, 1971. *A Theory of Justice*, Cambridge: Harvard University Press.

Regan, Tom, 1983. 'Paternalism, Freedom, Identity and Commitment', In *Paternalism*, ed. Rolf Sartorius, Minneanapolis: University of Minnesota Press: 113-118.

Scoccia, Danny, 1990. 'Paternalism and Respect for Autonomy', *Ethics*, 100 (2): 318-334.

Spellecy, Ryan, 2003. 'Reviving Ulysees Contracts', *Kennedy Institute of Ethics Journal*, 13 (4): 373-392.

Stone, Richard, 2009. *The Modern Law of Contract*, Hoboken: Taylor and Francis.

Taylor, Charles, 1985. 'Atomism', In *Philosophy and the Human Sciences: Philosophical Papers Vol. 2*. New York: Cambridge University Press: 187-210.

Valerius, Jukka, 2006. 'On Taylor on Autonomy and Informed Consent', *The Journal of Value Inquiry*, 40: 451-459.

VanDeVeer, Donald, 1986. *Paternalistic Intervention: The Moral Bounds on Benevolence*, New Jersey: Princeton University Press.

Waldron, Jeremy, 2005. 'Moral Autonomy and Personal Autonomy'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, Cambridge: Cambridge University Press: 307-329.

Wolf, Susan, 1989. 'Sanity and the Metaphysics of Responsibility' In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, New York: Oxford University Press: 137-151.

Young, Robert, 1989. Autonomy and the "Inner Self", In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, New York: Oxford University Press: 77-90.

Paper 3: Mental Competence and its Limitations

Abstract: In this paper I put forward my central claims with regard to the flaws of the orthodox approach in bioethics to psychiatric paternalism, and how this approach can be improved. In the first few sections, I build the case that our current model of mental competence, and the informed consent doctrine that utilises it, are flawed at a fundamental level. The deficiencies of the concept of mental competence have led to it being stretched into an ad hoc justification for paternalism, in which bias is almost impossible to prove in individual cases, yet appears embedded within broader institutional practice. I argue that this is not a product of professional malpractice, but rather the inevitable outcome of asking psychiatrists to adopt an impossible measure of whether involuntary treatment is justified. Mental competence is inescapably an evaluative concept that reflects our views as to the type of individual who we wish to include in our 'society of persons', i.e. within the mutual attribution of rational and moral agency.

I reach the conclusion that patient autonomy cannot be adequately understood using our current models of autonomy as a form of independent choice. Instead, we should understand autonomy in terms of the development and pursuit of an identity and concept of the good that is authentic to who one is. Autonomy, in this latter sense, is often a social project, reliant upon the existence of supportive relationships and environments. Similarly, *patient* autonomy is often best promoted through relationships of trust and care, rather than the attempt to enable independent decision-making.

Moreover, this paper demonstrates the tension between our aim of social inclusiveness in the attribution of moral agency, and our desire to impose standards best

approximating those required for effective moral agency. *None* of us are capable of objective moral agency, but we attribute it to each other in mutual recognition of our personhood. As first noted in the first paper of this dissertation, this demonstrates how the attribution of agency is inescapably a social and moral decision that is informed by medical science, rather than the value-neutral outcome of medical science.

Mental Competence and its Limitations

1. Introduction

1.1 Overview

The informed consent doctrine has near universal support within contemporary bioethics. Under that doctrine, mental competence is central to the permissibility of medical paternalism. A mentally competent patient is free from imposition of unwanted treatment: a doctor perturbed by a patient's refusal of treatment may urge the patient to reconsider and advise her of relevant information, but may not impose treatment against her will. In theory, this provides an absolute prioritisation of patient autonomy, whereby the mentally competent and informed patient has full authority over the pursuit of well-being through medicine. Recent philosophers have questioned whether the informed consent doctrine is sufficient to protect a robust conception of autonomy (O'Neill 2002, 23–27; Stirrat and Gill 2005, 130). The informed consent doctrine may, in practice, amount to little more than a choice to accept or refuse (with disastrous results) the sole treatment option provided by a medical institution in which patients are alienated by imbalances in power and information.

In this paper I argue that the informed consent doctrine is not simply insufficient, but relies on a flawed conception of mental competence that prioritises process over the substance of autonomy. I then argue that a more secure philosophical basis for involuntary treatment can be found in the protection of the patient's central goals and values. In practice, mental competence may remain a useful general indicator of whether involuntary treatment is permissible. However, in the context of psychiatric treatment, I identify occasions where paternalism should be imposed or withheld irrespective of the patient's mental competence.

1.2 Four Concepts

Competence is simply the state of being capable of performing a particular task effectively. The term's precision varies with the precision of the task that it refers to – for example, it is a far less precise description to call someone a competent swimmer than it is to say that someone is competent to swim the English Channel in clear weather. 'Mental competence' refers to the set of mental capacities that are required for a particular task. The term has gained a legal and medical significance associating it with decision-making and paternalism, but outside of that context one could also be mentally competent for other tasks such as passing an exam or solving complex mathematical problems. Our current interest in mental competence is as a guide to when paternalism is morally permissible. One might say, then, that we are concerned with the mental capacities required in order to have a moral claim to freedom from paternalistic intervention. No doubt the term 'mental competence' is used on occasion to mean precisely that. But such a notion of mental competence adds nothing to the analysis – it is just a round-about way of redescribing the question. It is more productive to talk about competence for the task of possessing some independent trait from which rights against paternalism can be derived. For this reason, in the context of psychiatric paternalism, 'mental competence' typically refers to one's capacity for meaningful autonomy, the foremost component of which is the capacity for rational agency.

In the previous paper, I discussed the liberal conceptions of autonomy leading to the appeal to *personal integrity* that underlies the liberal demand for restrictions upon medical paternalism. The appeal to personal integrity refers to our interest in living in a manner that is authentic with regard to our experience of self. Of the two

philosophical traditions that give rise to the appeal, I argued that the more secure grounding lies in a theory of liberal practical reason that can loosely be described as Kantian. However, most philosophical discussion of paternalism, especially during the height of philosophical interest in the subject through the 1980s, has rejected the Kantian approach in favour of a strict neo-Millian account (Feinberg 1984, 52–97; VanDeVeer 1980, 124–127; Kleinig 1983, 27–30; Arneson 1980, 470–489; R. Dworkin 1993, 242). Many philosophers use the term 'autonomy' to refer to any interest in authentic self-expression, referring to the Kant-Rawls account of practical reason as an account of 'rational autonomy' despite interpreting the account as one concerning rationality rather than the intrinsic value of self-authorship (Kleinig 1983, 26; Feinberg 1984, 96). Consistent with this, I will use this shared meaning of 'autonomy' to refer to the authentic self-expression that is the aim of the appeal to personal integrity, but without thereby preferring the neo-Millian account over the Kantian.

Our concept of *self* determines the scope of liberal autonomy and personal integrity. Our experience of identification and alienation shapes our personal identity and its authentic expression. For example, when coerced we 'choose' to comply and are a necessary part of the causal chain that leads to the choice, but we are alienated from our motivations, experiencing them as impositions rather than the authentic expression of our selves. These experiences are not restricted to conscious decision-making, and may apply to our involuntary movements, our relationships and the manner in which we are shaped by our environment (Meyers 2005, 29–31; Oshana 2005, 83–84). However, state paternalism involves intervention in deliberate choices, and in this context we are concerned with the autonomy of choices and actions, as determined by our identification with or alienation from our motivations and their outcomes. Liberal

philosophers have disagreed over the sense of self that ought be used to identify autonomy, as it relates to rights against paternalism. For example, Feinberg (1984, 112) and Arneson (1980, 474) imply a minimalistic sense of self, whereby we exercise autonomy by choosing in a manner consistent with the motivations that we immediately identify. This makes autonomy equivalent to voluntariness, and authenticity equivalent to intentionality. There is no substance to this sense of self beyond the person's immediate will, and so it is impossible for the content of a person's voluntary choice to undermine her self. Consequently, autonomy is a matter of process rather than content, whereby any decision can be an expression of autonomy so long as it is made with adequate information, mental capacities and an absence of impediments. In the previous paper, I argued that the appeal to personal integrity requires a more robust sense of self, whereby we posit a self that goes beyond our immediate experience and motivations. On this account, autonomy is defined in relation to the shaping of one's life in a manner that reflects one's more deeply held goals and values. This portrays autonomy as a matter of content as well as process, whereby a person's immediate and strongly held motivations may be at odds with the pursuit of autonomy by undermining the more central goals and values that characterise the way of living with which she identifies.

The *informed consent doctrine* is the implementation of the liberal approach to paternalism through the practice of medical ethics (Kirby 1983; Kihlbom 2008). Ordinarily, a doctor must obtain a patient's informed consent prior to any medical intervention. The exceptions to the doctrine involve process and capacity rather than the harmfulness or unreasonableness of the patient's choice: paternalism is permissible only where the patient lacks the capacities needed for informed consent, or where emergency circumstances make informed deliberation and consent impossible. In this

sense, the doctrine represents the high watermark of liberal rights against paternalism, adopting a concept of autonomy that is substantially equivalent to voluntariness. As such, the informed consent doctrine appears to presume that form of respect for autonomy as envisaged by Feinberg, where autonomy amounts to absolute personal sovereignty:

> *'respect for a person's autonomy is respect for his unfettered voluntary choice as the sole rightful determinant of his actions except where the interests of others need protection from him.'* (Feinberg 1984, p.68)

1.3 Liberalism and Mental Competence

Competence is not an intrinsically liberal concept, but it has become particularly useful to liberal accounts of paternalism. This is because the concept of mental competence appears to separate the criteria that *qualify* a person for liberal respect from the *content* of the person's choice (Buchanan and Brock 1989, 49–51, 65; Sandman 2004, 261–262; R. Dworkin 1986, 8. 13; Archard 1998, 21–22; Matthews 2000, 60–62; Faden and Beauchamp 1986, 266–267). The liberal on paternalism cannot tolerate paternalism that mandates an exclusive conception of the good life (Arneson 1980, 470; VanDeVeer 1986, 124–127; Gerald Dworkin 1971, 27–34; Kleinig 1983, 67; Raz 1986, 154–155, 204; Feinberg 1984, 3:54–68; Boddington 1998, 72–73). That is not to say that liberalism itself is morally neutral, but it seeks to preserve the permissibility of moral and cultural pluralism. Mental incompetence provides a criterion for paternalism that has the appearance of a descriptive concept (illusory, as I argue later in this paper).This allows the liberal to accept that paternalism is sometimes permissible without conceding state authority over the content of a person's self-oriented goals and

choices.

Since the end of the 1980s, liberal philosophers have become more sympathetic to the view that autonomy requires more than the possession of sufficient mental capacities together with access to information. We can possess the mental capacities required to make a particular choice with autonomy, yet fail in the execution of those capacities, due to weakness of will, social pressures, substance dependence or oppressive social structures. This more onerous concept of autonomy has passed into the theory of bioethics. Recent bioethicists have been more concerned with the lack of accessible, transparent and resourced treatment options than the imposition of unwanted treatment, characterising autonomy as arising from a therapeutic relationship based on care and trust (O'Neill 2002, 25–27; Bluhm 2009, 141–142; Taylor and Hawley 2006; Stirrat and Gill 2005, 130). However, medical paternalism remains governed by the legal and medical doctrine of informed consent. Even as they reason that informed consent does not necessarily amount to an effective exercise of autonomy, bioethicists do not apply the same critique to the authority of a patient's informed *refusal*.

At first glance, Mental health legislation forms a striking exception to the general dominance of the informed consent doctrine in both medical ethics and law (c/f Kirby 1983). The *Mental Health Act* (WA) 1996, *Mental Health Act* (NSW) 2007, *Mental Health Act* (UK) 1983 and *Florida Mental Health Act* 1971 provide broadly representative samples of such legislation in Australia, the UK and the USA, and each mandate the use of involuntary treatment where necessary to prevent mental illness from various statutorily defined harms to the patient, her finances or relationships, with no direct mention of mental incompetence or capacity to refuse consent. For example:

"26. Persons who should be involuntary patients

(1) A person should be an involuntary patient only if —

(a) the person has a mental illness requiring treatment; and

(b) the treatment can be provided through detention in an authorised hospital or through a community treatment order and is required to be so provided in order —

(i) to protect the health or safety of that person or any other person; or

(ii) to protect the person from self-inflicted harm of a kind described in subsection (2); or

(iii) to prevent the person doing serious damage to any property;

and

(c) the person has refused or, due to the nature of the mental illness, is unable to consent to the treatment; and

(d) the treatment cannot be adequately provided in a way that would involve less restriction of the freedom of choice and movement of the person than would result from the person being an involuntary patient.

(2) The kinds of self-inflicted harm from which a person may be protected by making the person an involuntary patient are —

(a) serious financial harm; and

(b) lasting or irreparable harm to any important personal relationship resulting from damage to the reputation of the person among those with whom the person has such relationships; and

(c) serious damage to the reputation of the person." (section 26, *Mental Health Act* (1996) WA)

Nonetheless, mental competence is embedded into the statutory definition of 'mental illness'. For example, section 4 of the *Mental Health Act* 1996 (WA) requires that, to be an illness, a condition 'must affect judgment or behavior to a significant extent'. The Florida legislation adopts a similar measure by requiring that to qualify as a mental illness, the impairment 'substantially interferes with a person's ability to meet the ordinary demands of living'. I note that the proposed *Mental Health Bill* (announced 16 December 2011, but currently a draft for public discussion) would make this role of mental competence explicit, via sections 12 and 25. This incorporation of concern for informed consent and mental competence is consistent with the treatment of involuntary hospitalisation programs by texts and studies operating from within a context of applied psychiatric ethics rather than moral theory (Beauchamp and Childress 2009, 110; Grisso and Appelbaum 1995).

My aim in this paper is to show that medical ethics remains deeply confused about the role of informed consent in medical paternalism, and that patient autonomy can be better protected by replacing the incompetence criterion for paternalism with a principle that medical treatment be consistent with a patient's goals and values. The informed consent doctrine, in its current form, fails to match medical paternalism as it is practiced and as it should be practiced. The result has been the conflation of evaluative judgments about the reasonableness of the patient's choice with the descriptive assessment of mental competence. Instead of separating the content of the patient's choice from the justifications for paternalism, the expanded concept of mental competence provides an illusion of normative neutrality that disguises what is

essentially an evaluation of whether the choice is reasonable.

The solution that I propose is to eliminate the gap between the content of the patient's decision and our grounds for paternalistically intervening. The moral authority of a patient's treatment instructions rests on something simpler than the competence model of informed consent indicates: *whether the patient's choice is reasonable with regard to the patient's own deep held goals and values.* We ought to judge patient autonomy expressly by reference to the content of the patient's choice, without the illusion of a purely descriptive justification, where that assessment is carried out with deference to the authority of the patient's own attribution of value. Instead of a right to choose independently of medical advice, a patient is owed respect for his or her personal values, such that medical treatment is not used as a vehicle for imposing a way of life that the patient does not identify with.

I intend to pursue this point through four parts. The first part provides an overview of the competence model of informed consent, as well as a more thorough statement of the problems immediately facing that model. In the second section, I review an example of how this manifests in the bias towards medical beneficence, wherein patients may be found incompetent to refuse treatment, but immediately competent to consent to that same treatment. My point is not to criticise the bias, but to further reveal that the competence model fails to meet its own aim of separating the content of patient decisions from the grounds for paternalism. In the third section, I argue that the normativity of mental competence cannot be reduced to medical misapplication, but is fundamental to the concept itself. The 'standard' of mental competence required for a decision does not reflect some universal fact of human decision-making, but is the answer to a *moral* and *cultural* question, that of how much responsibility we should

attribute to ordinary members of a society. In the fourth section I put forward my positive account of how the appeal to personal integrity, analysed in terms of its philosophical foundations in the previous paper, ought be practically implemented. I argue that our concern for mental competence ought be replaced with a concern for the substance of the patient's choices in light of her deep goals and values.

2. The changing shape of mental competence

2.1 The traditional approach to competence and autonomy

In the second half of the 20th century, there was a fairly strong consensus over the content of the capacities needed for mental competence. As a representative example, Buchanan and Brock (1989, 23–27) define mental competence as possession of:

- The capacity for understanding and communication of relevant information;
- The capacity for deliberation and reasoning; and
- A set of values or conception of the good, with sufficient internal consistency and stability in the values relative to the decision.

Buchanan and Brock intend these capacities to be interpreted initially in a threshold fashion, where sufficient capacity provides complete competence for the relevant decision. Where that threshold isn't reached the same capacities may be of interest in a comparative sense, to determine *how much* paternalistic oversight is best, keeping in mind that state paternalism and institutionalisation can bring their own negative effects (Buchanan and Brock 1989, 27–28).

The following decade, Grisso and Appelbaum (1995) summarised mental competence as arising from a 4-step procedure that is essentially the same as that given by Buchanan and Brock (splitting the first of Buchanan and Brock's criteria into two):

- The capacity to communicate a choice.
- The capacity to understand relevant information.
- The capacity to appreciate the situation and its likely consequences.
- The capacity to manipulate information rationally.

What both sets of authors leave unclear is the precise degree to which these capacities must be possessed to qualify for liberal limits upon paternalism. Putting aside for a moment the more specific debates over whether it is reasonable to require a higher standard for riskier decisions, or for refusal of treatment compared to consent to treatment, the only clear guide seems to be: (a) that there is some threshold level, such that the 'average' person is not rendered incompetent by being surrounded by Mensa members, and (b) the capacities required are far from absolute, i.e. we do not need perfect knowledge or infinite capacity to calculate paths of instrumental reasoning.

In their ever-influential *Principles of Biomedical Ethics*, Beauchamp and Childress (2009, 101) describe autonomous action as involving '*normal choosers* who act (1) intentionally, (2) with understanding, and (3) without controlling influences that determine their action...it needs only a substantial degree of understanding and freedom from constraint' [italics added]. Competence is associated with *normality*: understanding of the relevant information is also required, but the standard of competence is relativised. The implication is that we should expect competency to be available to those who are healthy members of the relevant culture.

2.2 The challenge of unreasonable attitudes and expanding the standards of competence.

The traditional mental competence framework concerns the mental capacities that are

strict pre-requisites for patient autonomy. As such, mental incompetence refers to lack of capacity or inadequate capacity, to be differentiated from a failure to effectively exercise those same mental capacities. In principle, the traditional formulation of mental competence does not place any requirements upon the content of a patient's choice. It refers only to descriptive features, being the patient's mental capacities.

It is vital to this conception of mental competence, that competence or incompetence can be judged independently of the content of the patient's choice. This is so even if we put aside the broader liberal interest in separating the content of a patient's choice from the justifications for paternalism. If the test of competence is mental capacity, then the patient who lacks adequate mental capacities should do so regardless of whether she accepts or refuses medical treatment. A grossly unreasonable decision might give us reason to inquire into the patient's mental state, but if our assessment must turn upon the content of the patient's choice, then we are judging something other than mental capacity.

The traditional framework of mental competence functions best when applied to individuals suffering from factual delusions, or from deficits in their ability to understand and rationally apply information in pursuit of their goals. If a patient is delusionally convinced that all around her are malicious conspirators trying to harm her, she is not simply choosing unreasonably: she is *incapable* of reasoning effectively regarding her medical treatment. As such, we can apply the mental competence framework in saying that medical paternalism is warranted because of the patient's incapacity, quite separately from the unreasonableness of the decision which results from that incapacity.

It has long been a fear within the neo-Millian line of liberalism that, due to the multiplicity of different views about how reasonableness should be determined, consideration of the *reasonableness* of a patient's decision will degenerate into an imposition of the common or dominant views of the community, or worse, the personal judgments of whichever doctor happens to be present (Buchanan and Brock 1989, 49–51, 65; Sandman 2004, 261–263; Faden and Beauchamp 1986, 266–267). The partitioning of the content and the authority of a patient's treatment choices is only possible if we can identify mental incompetence independently of the unreasonableness of patient's treatment choice. Once we move beyond the 'easy cases' of factual delusions, impaired comprehension and instrumental irrationality, our perceptions of unreasonableness and incompetence begin to converge. Mood and affective disorders, in particular, seem often to affect the sufferers' exercise of mental capacities in a manner that bears an ambiguous relationship with mental competence. Under the traditional mental competence framework, the autonomy-sapping effects of depression and other distortions of mood and affect are typically described in terms of impairment to the capacity for deliberation or the rational manipulation of information (Charland 2002, 44–45; Rudnick 2002, 152–153; Buchanan and Brock 1989, 65). Sometimes these impairments may be identified independently of the choice in question, bringing the patient back within the 'easy' category of cases discussed earlier – patients so affected by mania or depression that medical staff can identify them as too impaired to have moral ownership of a particular decision regardless of the actual content of the person's choice.

The cases become more difficult where an absence of mental capacities is inferred from the patient's choice itself, combined with some explanatory state that renders the patient vulnerable to mental incompetence. In 'The Inadequacy of Incompetence'

Culver and Gert (1990, 624–625) cite an example of a deeply depressed patient who has adequate understanding of all available information as well as insight into her condition, yet unreasonably refuses treatment. Buchanan and Brock (1989, 65) argue that incompetence is a preferable justification for intervention than the unreasonableness of the patient's choice, but this is to miss the point. The patient's depression only provides for the possibility of mental incompetence – it doesn't demonstrate that she is actually incompetent with regard to this particular choice. Mental illness – even noticeable mood disorder accompanied by an unreasonable refusal of treatment – is no guarantee of mental incompetence. If it were, the entire criterion of mental incompetence would be illusory; any mental illness combined with poor decision-making would serve the legal and moral functions that mental incompetence fulfills. Culver and Gert's example presents us with a possible cause of incompetence, a decision that would be explainable through incompetence, but the only evidence of incompetence is an increased susceptibility to mental incompetence combined with a significantly suboptimal choice. To infer incompetence in this manner is to stretch the concept so far that we deprive it of its justificatory power. Mental illness alone becomes enough to justify overriding a decision that medical staff view as unreasonable.

It is the distinction between the explanatory and the justificatory functions of mental incompetence that requires it to be identifiable independently of the choice itself. The presence of mental illness provides a useful explanation of why the patient declines to follow her doctor's recommended treatment – specifically, that the patient was rendered mentally incompetent. Whatever explanatory value may be available, depression here provides nothing in the way of a justificatory function. All we can do is *infer* that incompetence is one of two possibilities. The other possibility being, that

the person has competently decided not to have treatment – treatment, in the case being discussed by Culver and Gert, as well as Buchanan and Brock, consisting of Electro Convulsive Therapy, or 'ECT'. The side effects of ECT include sedation and short term memory loss, and it has been refused by those who have undertaken it in the past – it might not be as invasive a therapy as surgery or long-term sedative medication, but in a medical environment where some people take the utmost care over being subjected to blood tests, and where *any* impact upon memory, no matter how minor, might be considered a point of concern, it is not beyond credulity that a competent and informed individual might decline the option. In that scenario, the *justificatory* function of mental incompetence is not fulfilled.

The lesson from this is that a competence framework only *effectively* separates the content of the patient's choices from the justifications for paternalism if evidence of the relevant incapacity can actually be observed *independently* of the patient's choice. In recent years, however, some philosophers have sought to expand the scope of *observable* incompetence by specifying further capacities that can be practically measured independently of the patient's treatment choices. Kluge (2005, 297–298) is representative, adding emotional and valuational competence to the traditional, 'cognitive competence', framework. The relevance of incompetence isn't fundamentally altered: these capacities are still intended by Kluge to constitute the mental qualities required for the exercise of morally meaningful autonomy and personhood. But whereas the capacities to deliberate and to manipulate information cannot be externally observed save through the decision to which they apply, we can plausibly observe a person's emotional responses and, through inquiry, the values that she holds.

Certainly, an inability to form and identify with emotional responses, or an inability to

reach and be motivated by one's evaluative judgments, would prevent an individual from engaging in the kind of self-actualising autonomy that liberals seek to protect from paternalistic interference. When strictly confined to this sort of capacity judgment, an expanded competence framework may elucidate ways in mental illnesses observably deprive the sufferer of meaningful autonomy in ways other than factual delusions and instrumental reasoning. However, despite the usefulness of the extended framework when dealing with such broad inability to form appropriate emotional stimuli, this takes us no closer to resolving the kind of 'difficult case' discussed earlier in the context of mood disorders: those where the disorder provides a plausible explanation of why a choice may be mentally incompetent, but does not reveal whether any particular choice is made with incompetence and to what extent.

Unlike factual delusions, the relationship between incompetence and the mental impairment cannot be explained in terms of the logical requirements for reasoning: the patient is not deprived of the information or the capacity to apply it, but lacks the appropriate emotional and evaluative responses required to motivate her choice. Broad absences of these capacities might prevent a patient from forming motivations at all, or leave her apathetic with regard to the pursuit of her motivations, but a patient who has chosen to refuse treatment and demands that her choice be respected is not literally incapable of such motivation. Instead, in denying her mental competence, we would be claiming that she is not capable of the right 'kind' of motivation. But it is not at all clear what the right kind of motivation is, other than that some emotional states and evaluative leanings produce decisions that we believe are unreasonable.

It is informative that when Kluge provides an example of how the criteria may be applied, he justifies the patient's incompetence not by any description of how the

patient's impairment affects his exercise of irrationality, but by the inconsistency between the patient's values and the United Nations' declaration of human rights, apparently as evidence that the patient's values are grossly unreasonable (Kluge 2005, 297–298). If there is a distinction between the patient competently making the same decision, and the choice evidencing that the patient's impairment is sufficient to cause mental incompetence, Kluge treats it as irrelevant, and it is difficult to see how we could know which category such a patient would fall into. The extended competence framework leaves us in the same position as the traditional framework, deciding the patient's mental competence by the reasonableness of her choice.

By trying to place *all* involuntary treatment within the framework of mental competence, we invite a covert – and therefore unstructured – judgment about the reasonableness of the patient's treatment choices. The problem is that, despite its limitations, mental competence is not an empty concept; to deny a person mental competence is to deprive her of something important. Acknowledgment of another's mental competence requires us to recognise that she is relevantly similar to ourselves in the practices of giving, recognising and interpreting reasons. Our claim that her decisions are unreasonable must then be put in terms of an appeal to shared principles of rationality, rather than any implied or explicit assertion of inherently superior rationality. To deny a person's mental competence is to cut her out of the conversation about the rationality of her choice; it is to say that she is no longer capable of giving and recognising reasons for action concerning that particular choice. By inviting the gatekeepers of medical paternalism to infer incompetence from their evaluative judgment of the patient's treatment choice, we risk distorting the standards of mental competence, unfairly denying respect to members of minority cultures and to those who simply hold unusual but coherent values. This is disturbingly apparent in Kluge's

own elaboration upon his expanded competence framework. Elderly members of a minority culture are considered not merely wrong or unreasonable, but mentally *incompetent*, because their shared cultural view that life-extending treatment should be withheld once they are no longer economically productive members of their household clashes with the values of the dominant culture (Kluge 2005, 298). The problem is not Kluge's view that their cultural value or the elders that participate are unreasonable. Kluge may have a valid point to make about the value of life. The problem is that by finding his elderly targets to incompetent, Kluge excludes them from the moral conversation, ridding the paternalist of the need to justify his judgment in terms of an appeal to the targets' own values or to shared principles of rationality.

The challenge facing the mental competence framework is that by trying to bring all cases of involuntary treatment within the framework, we have distorted it such that incompetence is now routinely inferred from an unstructured judgment of unreasonableness, subjecting minority cultures to the values of the majority in the manner that liberal philosophers intended that the competence framework prevent. In the next two sections I argue that mental competence is a normative concept wherein we assert the responsibilities and capacities that a person *should* possess. By inviting the gatekeepers of medical paternalism to infer incompetence from unreasonableness, we impose a grossly illiberal standard for mental competence, wherein respect for personhood is confined to those whose values are considered reasonable.

3. Asymmetrical standards of competence

It has long been alleged that there is a tendency amongst medical professionals to only

question a patient's mental competence when the patient refuses treatment, requiring explanation and psychiatric examination in the event of such refusal whilst accepting a patient's consent to treatment with minimal scrutiny (DeMarco 2002, 232–233; Culver and Gert 1990, 624–625; Maclean 2000, 286). If a patient consents to treatment in accordance with the doctor's advice, no further inquiries into her competence are made. If the patient refuses to consent to implementation of the doctor's advice, then medical staff commence a thorough investigation into the patient's mental state, contact the patient's next of kin, and try to determine whether there is sufficient evidence of disorder for the patient to be found mentally incompetent. In part, this is just an attempt to gather more comprehensive evidence of the patient's mental state. With a greater potential harm riding on the decision whether to paternalistically intervene, medical practitioners might understandably devote more resources towards the assessment of mental competence. However, the contentious event is that the difference in treatment is not *just* one of requiring greater certainty of competence where the patient refuses treatment. A higher *standard* of competence is required for refusal of treatment, such that a person with mild or moderate mental impairment may be found, at the same point in time, competent to consent to treatment but incompetent to refuse.

Wilks (1999; 1997) and Buchanan and Brock (1989, 60) defend the practice as an appropriate response to *asymmetrical competence*, where different sides of the same choice require differing standards of competence On first glance, asymmetrical competence looks suspiciously like an authoritarian paternalism dressed in liberal terminology. A policy of making the standard of competence more onerous for decisions that are contrary to medical advice, seems to build a value of conformity, or at least of safety, into the measurement of competence.

To clarify, it is not problematic that medical staff are more likely to *inquire* into a person's mental competence in treatment refusals, even if this also has the practical effect of a more onerous standard for treatment refusal than for consent. DeMarco (2002, 241) distinguishes asymmetrical standards of competence from differences in the *evidence* that a particular standard is met. It may be that in a context of limited resources, the opportunity cost of comprehensively examining the patient's competence in all instances of consent is prohibitive. If so, a culture of more comprehensively inquiring into unusually risky treatment choices may be pragmatically justified. However, the differential treatment of consent and refusal goes beyond the evidentiary burden. As in Culver's and Gert's (1990, 160) early example, a patient who has already been found incompetent to refuse treatment may nonetheless be found competent when he changes his mind and consents, even though there has been no change in his mental state and even though he would be found incompetent were he to change his mind again and refuse treatment. Rudnick (2002, 152–153) describes with approval an identical case example. In these cases, the patient was recategorised from incompetent to competent with regard to the same decision, following a change from refusal to consent but without any alteration in mental state. The concern here can only involve the standard, rather than the evidence, of mental competence.

Wilks (1999, 156–157; 1997) defends the practice explicitly in terms of asymmetrical standards of competence, rather than different evidence of the same standard. He cites two analogies, neither of which are wholly satisfactory. In the first, he compares the relevance of physical competency in tight-rope walking, to that of mental competence more widely, overlooking that his choice of analogy – tight-rope walking – allows him

to posit an overriding goal of safety shared by paternalist and participant alike, thus averting the entire basis of paternalistic dispute. His second example is that of competence in the use of the stock-market, again presupposing an overriding aim – this time financial security – through the parameters of his example. The source of conflict between the protection of patient autonomy and medical paternalism is that the ultimate goals of patient and medical staff may clash. Wilks' examples have something relevant to say, but it is about autonomy and self-authorship, not competence. They illustrate our identification with more central plans and ways of living, sometimes ahead of our immediate operative desires.

Wilks (1999; 1997) appeals to our assessment of *physical* competence for dangerous tasks:

> *'The same task performed in different contexts can occasion different levels of risk, and therefore require different levels of reliability in the doer of the task, and therefore different levels of competence.' (Wilks 1999, 156)*

Wilks does *not* mean that the greater potential for harm makes a task any more difficult. His claim is that even though the capacities required for a fixed level of competence are the same, the *standard of competence* required is greater. To illustrate this, Wilks (1997, 419–420; 1999, 156) asks the reader to consider a person's competence to cross a high-wire tightrope, and then compare it to that person's competence to cross that same tightrope once the safety net is removed. The response that Wilks expects, and I expect that he is correct in this, is that the change in the potential harm alone is sufficient to require that the performer has a greater level of skill and experience before proceeding.

Wilks' tightrope example concerns the skill required to *avoid* a negative outcome, whereas greater mental competence only affects our understanding of the outcome from pursuing a particular option. The example relies on analogy rather than directly illustrating Wilks' claim and there is scope for uncertainty about whether physical competence and mental competence are relevantly similar enough to make the analogy work (Checkland 2001). As Wilks (1999, 157) acknowledges, in the high-wire example a powerful value of safety is built into the person's idea of 'wellness'. The person wants (or would want, if choosing consistently with her central values) not simply to cross the high-wire, but to do so safely – and physical competence makes the same action safer. Safety is presumed to be more deeply held by the tightrope walker than the goal of merely crossing the wire. If the walker was deliberately and soberly reckless regarding his safety – say, deliberately putting his life in grave danger as part of a protest against a horrendous occupation – Wilks' example no longer provides clear support for his contention.

Wilks' (1997, 421) second example is more interesting in its involvement of decision-making capacities, but is still of dubious relevance to mental competence as it applies to paternalism. This time Wilks refers to his own understanding of the stock market, and claims that if presented with an offer to purchase speculative shares that he knows little about, he would be unable to judge the likely profit or loss, and hence could not feel competent to make the purchase, but as his opportunity cost in forgoing potential profits is rather more trivial he remains competent to refuse the offer. Even putting aside the potential artificiality of supposing a shared ultimate goal, Wilks' treatment of the example is groslly inconsistent with actual practice. Wilks doesn't describe his account as revisionary (and his belief that he matches popular intuition would suggest

the opposite (1999, 155), yet this is not at all how competence to engage in financial transactions actually works. The absence of severe mental impairment, is sufficient for legal control over one's own financial matters[1] Certainly, Wilks might be considered foolish if he forgoes seeking information or advice, but his uncertainty and limited expertise would not place him in any danger of the state stepping in and transferring control over his finances to an administrator or financial guardian to act in his interests – which is *exactly* the kind of institutional usurping of legal rights and physical liberty that Wilks is seeking to justify by analogy between financial and medical competence. Wilks is confusing the ability to determine whether the investment would be profitable (to do the activity well) with the ability to decide whether he should risk the investment in light of his goals and values (to decide competently). As Wicclair notes (1999, 152-153), Wilks (unlike the children and mentally impaired individuals who we *actually* find incompetent) possesses the second-order desires, capacity for reflection and the ability to recognise his lack of knowledge, such that he can assess his lack of first-order skills and place the resulting risk in the context of its relevance to his deeply held goals and values.

Both Wicclair (1999, 151) and Cale (1999, 138–139) argue that by incorporating risk into the measurement of competence, we impose a normative judgment about the relative value of safety. Wilks, acknowledging this normative imposition, responds :

> *'What is the normative standard we employ in way of understanding what good medical decision-making is?... patients should make treatment decisions that are reasoned products of their own well-entrenched personal values. From this*

1 A typical Australian statute governing the paternalistic supervision of financial self-governance is the *Guardianship and Administration Act 1996* (W.A.). See sections 3 (definition of 'mental disability'), 4 and 64, especially the presumption of competence under 4(3) and the need for a mental disability before an administration order can be made under 64).

> *point of view, patients are more competent to make treatment decisions insofar as they are more able to achieve this goal.' (Wilks 1999, 157)*

Wilks, like his critics, believes that the standard of competence ought follow the person's own attribution of value. Wilks' account diverges from Cale's and Wicclair's in that he makes a sweeping assumption about the content of the 'well-entrenched personal values' above. Wilks assumes that, in all cases, a patient's most central value, relevant to her treatment choice, must be the preservation of health. This doesn't mean that the patient subjectively values her health directly. Rather, enough of her deeply held goals rely on her health, directly or indirectly, such that it is at the center of her web of interconnected wants. I argue in the next paper that respect for patient autonomy should not preclude a bias towards health nor towards the preservation of patient life. However, to posit such a goal independently of the patient's own attribution of value excludes some of the more famous disputes in the ethics and law of medical paternalism: the clashes between medical values and the commitments of religion and ideology, as found in the Jehovah's Witness' prohibition on blood transfusions, as well as the pursuit of dignity and self-determination as motivations for euthanasia.

Of course, most of us do place an extraordinary value upon our own health. We are necessarily *embodied* agents and we cannot wholly separate concern for our health from self-concern. We ordinarily *presume* that others have good reason to be concerned about their well-being, unless we have evidence to the contrary. This could quite plausibly justify a greater cautiousness regarding choices that appear to harm a patient's health, manifesting in a demand for greater evidence that the patient's choice has arisen from an authentic exercise of her autonomy.

s

But the bias against refusal of consent, and Wilks' account of asymmetrical autonomy in defence of it, goes further than a demand for greater evidence before committing to a decision. There is a difference between requiring further inquiry, and imposing different standards of proof after that further inquiry has been carried out. Wilks explicitly equates asymmetrical reliability of evidence with asymmetrical standards of competence through the example of an exam, whereby increasing the difficulty of the exam both raises the standard required for a pass, and thereby raises the reliability of the evidence that those who pass have an adequate knowledge of the subject (1999, 155). Even within that example, the equivocation is flawed. An exam can be more difficult without thereby being more reliable: the exam could consist of only one question (analogous to the cursory examination of patient competence prior to refusing treatment), or it may suffer procedural flaws such as undetected plagiarism. Similarly, one could improve the reliability of evidence without raising the standard, by increasing the *number* of assessments while keeping the difficulty constant. The bias against refusal of treatment is problematic because it continues to assert an asymmetrical standard of competence *after* the further inquiry warranted by a refusal of treatment has been carried out. In requiring a greater *standard* of competence, medical institutions are not investigating more clearly the patient's own goals and values, but rather imposing a value (that well-being should be prioritised) upon her.

Through the asymmetrical standard of competence, medical institutions determine the patient's incompetence by reference to the content of her choice. Moreover, here the conflation of competence and content cannot be explained in terms of utilising the content of the patient's choice to assess her patient's mental state. Even if we dismiss the practice as an erroneous application of the concept of competence – which it

almost certainly is – the more troubling matter is that the bias is made possible by the absence of any 'objective' standards against which the capacities needed for competence can be measured. The bias is made possible precisely because in our ordinary application of the concept of mental competence, the patient's competence or incompetence is routinely derived from the content of her decision. It is significant that the examples that philosophers have used to illustrate asymmetrical competence have been mood disorders that present a potential, but unprovable, cause of the patient's refusal of treatment. These are the same kinds of case as those discussed in the previous section as requiring the conflation of competence and the content of the patient's choice. The bias against refusal to consent illustrates our inability to determine in such cases whether or not the patient's apparent unreasonableness is the result of mental incompetence.

4. Competence as value-driven

Bioethics has often discussed standards of mental competence as though they are descriptive facts that we discover about the world and ourselves. Competence gains its practical importance from normative concepts but is itself usually treated as a descriptive concept: measuring the mental capacities needed to achieve a certain level of processing, with normative concepts and ideologies establishing our interest in whether that level of processing can be achieved, and what rights or obligations ought be attributed as a result. In exploring the ways that mental illness may deprive a patient of mental competence, it is common to assume that we have a shared, fixed conception of the standard of competence needed for personhood and moral responsibility, one that we can ascertain from a proper understanding of the patient's choice and the meaning of autonomy (Grisso and Appelbaum 1995, 106–109; DeMarco 2002, 232;

Wilks 1999, 155–156; Rudnick 2002, 151–153; Buchanan and Brock 1989, 65; Stanley and Stanley 1982, 2–3; Charland 2002, 137–138). Under this approach, we 'discover' that a mental illness renders an individual incompetent, by finding that it reduces the patient's capacity below the standard required for a decision of that risk and complexity. In this section, I argue that the finding of mental incompetence is an evaluative *decision*, rather than a descriptive discovery. There are also some types of choice (say, complex investment decisions such as in derivative security markets) where we typically allow people to bear full responsibility for their decisions on the basis of what seems more like a matter of economic and social policy than a claim to meaningful autonomy. In recent decades, the hidden evaluative nature of 'mental competence' may also have allowed a gradual creeping upwards of the standards required for mental competence: as we have become more aware of social and biological influences upon our decision-making, we may have increased the standards we require for moral autonomy and responsibility. If so, this necessarily requires a reduction in 'respect', i.e. in the proportion of people who we view as mentally qualified to participate in the moral conversation about how they should live.

There is a moral and cultural danger to this. By raising the standards for mental competence, we diminish our societal capacity for the tolerant co-existence of different cultures. Cultural differences are not merely local adaptations or cosmetic variations on the pursuit of the same conception of the good; the Jehovah's Witness' refusal of blood transfusions, and the welcoming of passive euthanasia discussed by Kluge, reflect deep disagreement about the reasonableness of their cultural values. Cultural pluralism requires that we tolerate such disagreement and that we settle disputes about the right way of living by appealing to each others' capacity for reason. A heightened standard of competence risks denying respect to cultural minorities, such

that highly personal decisions are made subject to the values of the dominant culture.

Mental competence, as we currently apply it, is a feature of normalcy. Most of us will never experience psychiatric paternalism, nor any other sustained paternalism arising from an allegation of mental incompetence. We are accustomed to thinking about mental incompetence as a label applied to those less capable than us, rather than a label that is potentially applicable to us. The 'average person' is often assumed to be lacking in knowledge, and susceptible to occasional weakness of will, and consequently there is a myriad of fairly uncontroversial legal regulation designed to ensure that people receive expert advice before acting upon decisions that require specialist knowledge, and that people don't suffer catastrophic harm due to momentary carelessness (e.g. seat-belt requirements). In addition, some seemingly personal activities (e.g. drug and alcohol use), are viewed, rightly or wrongly, as social vices, subject to regulation. These forms of broad paternalism imply nothing about our mental competence. They are exceptions to the liberal framework of mental competence altogether; restrictions imposed *despite* our mental competence. Whenever paternalism is predicated upon an estimation of an individual's personal mental capacities, such as in overriding a medical patient's refusal of treatment, or denying someone the authority to enter an otherwise legal contractual agreement, normalcy is taken as sufficient for mental competence.

What I mean by the 'average person' is simply an adult with normal mental function. To the extent that mental competence involves temperament or values, the average person holds the temperament and values that most members of her society find appropriate to her circumstances. In this sense, 'average' means ordinariness or normalcy of function, and is not necessarily representative of the disorders and

capacities held by most people. Given the natural variation in mental capacities and the prevalence of mental illnesses such as phobias, depression and anxiety, it may well be that most people suffer from a small number of minor mental disorders during their lifetime (very few real people are *uniformly* average).

The competence of the average person has been a mostly silent assumption. Despite their extensive analysis of the kind of capacities relevant to mental competence, even the highly influential works on bioethics by Buchanan and Brock (1986, 23–27) and Beauchamp and Childress (2009, 101) have given very little guidance as to the *standard* to which these capacities must be held. Rather, it seems that as a matter of presumption, normalcy *is* the standard against which incompetence is determined. Beauchamp and Childress (2009, 101) state this explicitly by defining mental competence in terms of 'normal choosers', while it is broadly taken for granted that ordinary mental function is sufficient for mental competence (Rudnick 2002, 153; Wikler 1983, 84).

This confidence in the competence of the average person is not based in any assumption of infallibility. In empirical studies of patient competence, even the mentally health control groups display a generally poor comprehension of the information relevant to consent (Grisso and Appelbaum 1995, 110–116; Stanley and Stanley 1982, 2). In commercial contexts, we sometimes go so far as to *expect* that differences in the capacity to comprehend and apply information will be used to gain an advantage over other less capable decision-makers. Yet participants in complex investment markets are not required to be aware of, nor able to comprehend, influences upon the market value of investments, nor to have more than a general appreciation that such investments involve risk (as opposed to a knowledge of the actual probability

of loss given the market situation).

This sets up an apparent distinction in the significance that we attribute to differences in mental capacity: the mentally impaired are considered less competent than the average person, but no such distinction is made between the average person and someone with exceptional mental capacities (whether through greater capacity for comprehension, or by being less susceptible to lapses like weakness of will). This has sometimes been explained in terms of diminishing returns from increased mental capacity, such that increases beyond the capacity of the average person provide little or no further benefit to autonomy (Buchanan and Brock 1989, 27; Wikler 1983, 89). This may be true for some choices, but there is a danger that we take for granted the losses of 'ordinary' limitations, given the potential to further minimise losses from conditions such as skin or lung cancer. Whilst a greater understanding of medicine or finance won't eliminate risk from these choices, it may enable the decision-maker to reach a better understanding of the nature of the risk, a closer estimate of the likely outcome, and thus greater authorship through her choice. Yet the average person's relative disadvantage is not taken to be a sign of mental incompetence.

There are two kinds of story that we could offer as explanation of this assumption of broad competence. Firstly, and most commonly, we might say that although the average person isn't perfectly capable of self-authorship, distinction, perhaps we can draw a line that marks the rough centre of the middle ground between unambiguous competence and unambiguous incompetence, and that either by some good fortune, or evolutionary or social design, it happens to match the capacities of the average person. The other kind of story is to say that the average person is broadly mentally competent because we decide that ordinariness should be the standard at which people should be

treated *as though* they are capable of exercising meaningful personhood. We decide *first* the types of people to whom we should attribute the moral responsibility and autonomy of personhood, and *then* establish a standard of competence by reference to their capacities. In a particularly cynical variant of this story, we might say that normalcy implies mental competence only because rule by majority encourages us to set a standard whereby most of us have our sense of personal agency authenticated. The common element to this kind of story is that the requirements for mental competence are determined by an evaluative judgment about the kind of society we want, and the values that should guide the relationship between society and individual.

For 1980s neo-Millians like Feinberg (1984, 52–97), the model of liberal autonomy was that of a person capable of forging her own identity independently of her social and cultural environment, other than that required in her formative years for the development of her mental capacities. Sociology's actuarial studies have shattered this portrayal of persons as causally free choosers, by recasting our voluntary life choices as the statistically probable consequences of unchosen socio-economic factors (Kemshall et al. 2006; MacDonald 2006; Rigakos 1999). This is not to say that we are governed by broad descriptors of class or race, and sociologists have at times despaired at the misinterpretation of actuarial studies to exaggerate the influence of such broad descriptors over individual differences (Cullen et al. 1997). Instead, we are fundamentally social beings, influenced by cultural and social influences we can't entirely escape, subject to gross power imbalances in the availability of information and opportunity, and living in an environment of such technological and bureaucratic complexity that we are routinely reliant upon formal relationships of care and advice. Whilst many of us hopefully have cause to reflect upon our values at some stage of life, our values do not originate by our reasoning our own will into existence, complete

with motivations and goals for which we are entirely responsible, and our moral and cultural similarity to those with whom we share geographical proximity and common state membership (sometimes bearing the influence of the latter without even having the former) ought defeat any notions of the innate liberal self-created being. Further to this, the specialisation of knowledge and advancement of medical procedure means that we may face decisions that are vital to the pursuit of our goals and values, where a personal understanding of the relevant information would require an extraordinary depth of knowledge and considerable intellect. When faced with a decision whether to pursue complex surgery, or to choose between different treatment paths, for most of us the practical extent of our capacity for self-authorship is little more than to decide whether to trust or to reject our doctors' recommendations. The liberal pursuit today, insofar as it has any relevance to our philosophy, is not a declaration of one's independent self to be imposed onto the world, but one of finding identity and resisting self-alienation in a world where the average person is vividly aware of the external impositions upon our would-be persona. The medical arena is one area where liberalism remains the undisputed centre of ethical debate, whereby no amount of recognition of our interconnectedness can shake the firm conviction that the forcible imposition of unwanted medication is too great a violation to be permissible without some explicitly liberal justification. To pursue liberal autonomy in this context is to mould an identity and narrative from our relationships and cultural context, finding autonomy *in* these connections rather than our freedom from them (Meyers 2005, 37–38; O'Neill 2002, 83–94; Friedman 2005; Oshana 2005, 84–94). But this strips the presumption of competence of its foundations. If the human capacity for autonomy is not that of independent self-determination, but of shaping ourselves through relationships and institutions in which we are disadvantaged through asymmetric distribution of power and information, then why should we assume that the average

person is broadly mentally competent?

.

In one of the few pieces directly addressing this matter, Wikler (1983) tries to explain the presumption of broad competence as a product of social construction:

> *'[the] threshold of competence in our society falls at or just below average because, first, the level of difficulty involved in key life tasks is in large socially determined; and, second, because society stands to gain by setting this level [i.e. the complexity of the task] so as to render the average person competent.'* (Wikler 1983, 89)

Wikler acknowledges the exceptionalism inherent in our distinguishing 'normal' limitations upon self-determination from mental impairment, but seeks to preserve the common depiction of mental competence as an objective feature of the relationship between the world and ourselves. The difficulty is that this explanation exaggerates the calculation and cooperatitive intention with which our environment is socially constructed. Many social and financial tasks are designed not by an individual or group with their own or society's interests in mind, but instead by the complex and unpredictable convergence of many independent operators all pursuing their own goals. Most industrial and investment markets take such a form. Whilst the forms, rules and procedures for participating in the share market are designed with the intent that most people can participate, the actual behaviour of the market itself is designed by no-one. Furthermore, social participation is no guarantee of broad self-authorship, and people matching Wikler's description of the 'average person' have often been subjugated into disempowering social roles due to their race, gender or social class. Even if we accept that democratic or economic forces favour making key social

institutions broadly accessible, it is difficult to see why this should imply a broad capacity to exercise meaningful self-authorship through those institutions.

We do not *discover* the broad mental competence of the average person, we *declare* it. None of us are capable of absolute self-authorship. Our responses to gender, race and sexuality are embedded into the perspective from which we experience the world (Oshana 2005), our identity is shaped partly by the opportunities we have available to us, and we are routinely subject to weakness of will, socially inbued prejudice and emotional impediments to deliberation. While some of us might be more prone to self-reflection, the perspective from which we undertake such reflection is itself a product of biological and environmental causes that are external to our sense of self. If competence was based on relative capacity, then only the extraordinarily capable would be competent. Instead, the average person has broad mental competence because we judge that we should act *as though* those of average mental capacities ought take responsibility for their actions and be attributed moral autonomy.

The relation between the imperfect capacity for autonomy required for mental competence and the moral autonomy it grants us invites comparison to the Kantian theoretical and practical standpoints. Through our ability to study our physical and social environments, we know that we are inescapably part of a universe governed by physical causality, and that the perspective from which we perceive and reflect upon ourselves and our environment is the product of biological and environmental influences. Yet despite this awareness, we experience the world from a perspective of agency and it is this practical standpoint that is relevant to our identity as persons (Hill 1989, 97–98). Similarly, we can recognise that the capacities that comprise mental competence fall short of allowing a person to exercise unfettered choice over her

character, while attributing her with full moral ownership over her character and the choices that reflect her character. The two sets of concepts are also related in a more direct way: the practical experience of agency is not restricted to those of average or greater capacity for autonomy, but we only validate the practical standpoint of those who are mentally competent. Insofar as an individual is mentally incompetent, we insist on addressing her as caused rather than willed, such as by mental illness or compulsion.

Most importantly, the standard that we set for mental competence determines *whose* values and principles we take to be relevant to determining rationality. By acknowledging others' personhood, we recognise that they are capable of giving and applying reasons, such that any claim that their actions are unreasonable must be mutually intelligible (discussed in the second paper of this dissertation: 'Reasons, autonomy and paternalism'). As such, we recognise that a person's own goals and values, to which she attributes rational authority, are sources of rational motivation for action. By labeling a person mentally competent, we legitimise her goals and values as sources of reason. Her goals are not always morally correct or legally permissible, as other peoples' interests may provide good reason for preventing an individual from pursuing a harmful goal, but they may rationally motivate an individual, subject to any more central or deeply held commitments.

In summary, mental competence provides a measure of the extent to which an individual is allowed to *participate* as a member of society, rather than as society's ward. If we are mentally competent, our social role is characterised in terms of our practical experience of agency, whereby our identification with goals and values renders them meaningful as sources of reason for action. To be labeled mentally

incompetent is to have one's perception of agency rejected, such that instead of participating in societal institutions, one is *acted upon*. Societal institutions may impose reasons for action upon the mentally incompetent, without appealing to the individual's own principles and values. For example, in paternalistically imposing psychiatric treatment onto an individual who is mentally incompetent to refuse treatment, the state asserts the patient's health as a source of reason for action, without needing to address the patient's subjective motivation for refusal so long as the patient is incompetent with regard to the forming and application of that motivation to the treatment choice. The boundaries of societal participation and membership is something that we assert as a normative judgment, not something we discover amongst the objective features of the world. As a determinant of societal participation (and thus membership), mental competence is a normative concept establishing the kind of society that we want to have. It gains its significance not as a measure of the objective traits of an individual, but as an assertion of who should be included within our society, and under what terms.

It is in this context that the traditional mental competence framework, with its focus on the possession of ordinary mental capacities, presumes a liberal construction of society. In Edwards (2009, 80–84) I explained that our susceptibility to ordinary impediments to autonomy – dysfunctions arising from heightened emotion, weakness of will or ordinary limitations in our capacity to understand and apply information – are part of our character. Through our shared concept of personhood, we are socially identified with these limitations (regardless of whether we subjectively identify with them) and expected to take responsibility for them. The traditional mental competence framework is concerned only with those traits that *aren't* part of who we are, i.e. dysfunctions that concern our mental or neurological health by virtue of being *external*

to our character. The framework is inherently liberal through its aim of separating mental competence from the content of a patient's choice *and* her character – it asserts that participation in society ought not be restricted by *who* a person is or the conception of the good life with which she identifies.

So rather than reflecting our objective capacity for self-determination, the broad competence of the average person arises because we assert that: (1) participation in society, i.e. personhood, should not be restricted by reason of who one is, and (2) our ordinary susceptibility to impediments to self-authorship are part of who we are, rather than external impositions upon one's self, and thus do not restrict personhood. The significance of this is that we can shift the standard of the capacities required for mental competence qualitatively and quantitatively, and that by doing so we alter the principles governing social inclusion. For this reason, an expansion of the requirements for mental competence to justify increased paternalism can lead to some unintended consequences. When we use mental competence to mandate particular values, or attitudes towards deliberation, we make these things into requirements for social inclusion (in the context of that decision). This isn't a merely theoretical concern. It comes with the threat of paternalistic enforcement, overriding the person's self-oriented choice and possibly her personal or cultural values.

This might be exactly what we intend to do – we may find a particular value or decision so alien to the kind of society we could endorse, that we refuse to acknowledge any claim to it providing a reason for action. Nonetheless, we ought to recognise that to do this is distinctly illiberal, and I suspect it is far more illiberal than many advocates of the competency model would be comfortable with. By raising the standard of competence to exclude particular value judgments or goals, we deny the

need to actually demonstrate that the person's choice is unreasonable, by delegitimising the attribution of reason that motivates the choice. The effect is to exclude people and possibly cultures, contrary to the commitment to the cultural tolerance and pluralism that the competence model is intended to protect. The concept of mental competence itself needs correction. If we are going to impose an evaluative judgment through the concept of competence, we ought to clarify *whose* values provide the basis for this judgment. This cannot be achieved through a descriptive account of mental competence, that denies that the gatekeepers of medical paternalism are making such judgments.

By having a concept of mental illness and impairment that incorporates evaluative judgments, and an account of mental competence that denies the need for them, the decision to impose psychiatric paternalism loses both transparency and fairness. I noted in the last section that the apparent lack of any objective measure of the capacities needed for competence makes it possible for systems of medical paternalism to be broadly biased against patients who refuse treatment, without any means of verifying the presence or absence of bias in individual cases. In this section I hope to have gone some way to explaining why: there *is* no normatively neutral means of assessing mental competence, and by overlooking this, the dominant account of mental competence doesn't stipulate the principles by which the evaluative judgment should be carried out. Without a clear stipulation of whose values should provide the basis for any evaluation of the patient's reasonableness, the assessment of competence lacks accountability and it becomes permissible for doctors to apply their own values, or the values of the medical institution or the state, contrary to the patient's own attribution of value.

5. Reason-driven paternalism

5.1 The Limitations of Mental Competence

In the past three sections, I have sought to illustrate how the common application of the informed consent doctrine, with its focus on mental incompetence as a necessary requirement for medical paternalism, undermines its own purpose. Mental competence is an unavoidably evaluative concept, that in its present form serves to disguise, rather than police, the imposition of goals and values foreign to the patient. The purpose of the work so far has been to build a case for revising the informed consent doctrine.

I note in advance that I leave an important part of this account for a later paper. Patients' goals and values change, especially when they are tested through times of great stress. We need some means of determining whether a change in a patient's values should be considered an exercise of autonomy, or as a barrier to autonomy imposed by illness. In 'Beyond Mental Competence' (Edwards 2010), I endeavour to provide that account. For now, I put the case for a revised understanding of competence and autonomy into which this account can fit.

In the latter parts of the previous paper, I argued that autonomy can be best understood in terms of the pursuit of *personal integrity*, i.e. the pursuit of the goals, values and other connative states that are most authentic to the patient as a person. This view of autonomy concerns our pursuit ofa way of living with which we identify, rather than a specifically *individualistic* pursuit. It appears to place very little importance upon individual atomistic choices, except where the patient holds such choices dear to her identity. In the medical context, this revised conception of competence, in the form of *personal integrity*, quite often requires the assitstance of medical expertise or even an entire medical team. It does not rule out the possibility of insisting upon individualism

per se, but for the great many of us who, in the medical context, are concerned more with outcomes and relationships of care and trust rather than battling our condition from a position of strict independence, it provides a manner in which autonomy can be made consistent with a shared project, while still laying out demands that such a project must meet (from the goals it must pursue and values it must be characterised by to the nature of the relationships that must comprise it).

Most of the time, this change in emphasis would make no practical difference. However, there are common occurrences where the singular focus upon mental capacity leads to absurdity. One tragically common example is where a patient suffering from chronic and severe mental illness (such as schizophrenia) is still mentally competent, but has begun a medically preventable mental deterioration as a result of ceasing medication. While some such patients no doubt have come to a considered preference for the symptoms of untreated illness over the side-effects of medication, many cease treatment for reasons that are substantially less than a deliberate exercise of autonomy, such as a lack of insight into their condition, or a illness-induced disorganised psychological state. Many of these patients end up on what might colloquially be termed a 'hospitalisation revolving door': entering treatment only when their mental state has deteriorated to the point of total mental incompetence, eventually being discharged from their involuntary treatment status once their competence has returned, only to cease treatment and restart the cycle of mental deterioration and hospitalisation (Kress 2006; Botha et al. 2010; Kastrup 1987; Spellecy 2003). The cycle makes meaningful autonomy deeply improbable – even during periods of competency, the patient lacks the sustained stability needed to reform relationships, find stable and independent accommodation and pursue the goals and commitments that they aspire to. To refrain from intervention where deterioration

is a near certainty, and will be utterly disastrous to the patient's preferred way of living, is to prioritise the process of competence over the substance of autonomy.

A second example is the legislative disempowering of psychiatric living wills, whereby an informed, competent and considered choice to forgo some or all medication in the event of mental illness can be overriden by a psychiatrist once the patient becomes mentally incompetent. By again making the permissibility of psychiatric paternalism a matter purely of assessing the patient's competence, her substantial exercise of autonomy is rendered meaningless. Whilst there may be some public interest in ensuring that the severely mentally ill receive adequate treatment, such involuntary treatment is more commonly imposed paternalistically, being triggered by concern for the patient's well-being rather than any criminal or civil dangerousness. In both of these examples, the better option would be to consider the patient's central values, and how she can be assisted in a manner consistent with those values.

5.2 Standards of autonomy and the appeal to personal integrity

In the initial sections of this paper, I have built a case that the informed consent doctrine encourages a hidden conflation of the patients' competence and the content of the patients' decision. This conflation reverses the protection that the competence account is intended to bring. Instead of applying paternalism strictly in relation to patients' mental capacity, we restrict personhood so as to permit intervention in choices that we find grossly unreasonable. More perniciously, by labeling the patient incompetent, we infer from her perceived unreasonableness that she is not capable of rational decision-making. By taking deviance from social norms to be evidence of incapacity, we impose those norms as constraints upon self-authorship.

We might respond to these concerns by endeavouring to apply the competence model 'properly'. For example, we might restrict the competence framework to consider only those mental capacities that can be measured independently of the content of the patient's choice. However, this does nothing to resolve the conflict between our theoretical commitment to the separation of competence and the content of the patient's choice, and our firm and widespread conviction that there is something less than fully autonomous about the irrational decision-making often displayed by patients suffering from mental illnesses such as mood disorders, even though they may retain an adequate understanding of the information relevant to their decision (Culver and Gert 1990; Buchanan and Brock 1989, 65; Rudnick 2002). It is unlikely that any bioethicist, outside of those few that remain opposed to involuntary psychiatric treatment altogether, would view the complete exclusion of such patients from psychiatric paternalism as a satisfactory outcome.

Personal integrity refers to our interest in pursuing a way of living that is an authentic expression of our self. The appeal to personal integrity is the claim that the 'good life' for a person is one lived in accordance with one's personal integrity, even at the cost of our well-being. The appeal has a basis in liberal accounts of autonomy and practical reason, whereby the self is the ultimate authority over our self-oriented interests and our reasons for action respectively (as discussed in the previous paper). A representative passage is provided by Ronald Dworkin (1993, 242):

> *'Recognizing an individual right of autonomy makes self-creation possible. It allows each of us to be responsible for shaping our lives according to our own coherent or incoherent – but, in any case, distinctive – personality. It allows us*

> *to lead our own lives rather than be led along them, so that each of us can be, to the extent a scheme of rights can make this possible, what we have made of ourselves.'*

The appeal to personal integrity requires filling out, in terms of what kinds of choices fall within concepts like self-creation, self-authorship and authenticity. In moving from a general statement of principle to a practical system of rights, we need to know how to unpack what it means to authentically express one's self. In the latter part of the last paper, I began this task by noting that autonomy could only hold the kind of value that the appeal to personal integrity suggests, if we view it as a matter of shaping or designing one's way of life, rather than just exercising control over individual atomistic choices. That is, the 'self' relevant to autonomy and morality is a persistent one. A persistent sense of self, of course, is assumed within most of our interpersonal relationships, as well as our systems of morality and justice. The concept of friendship, in particular, requires that we persist not only in terms of remaining the same fundamental 'being', but that we remain the same character, with a continuity of commitments, attitudes and values. Friendship implies mutual (albeit imperfect) knowledge of each other *as persons*, beyond just our immediate operative desires. This requires that we have substantive qualities – a persistent *character* or personality, which retains responsibility for past actions and that can make commitments that will morally bind us in the future. In 'Beyond Mental Competence' (Edwards 2010), I provide an account of this temporally persistent, but sometimes severable, self.

The competence approach to paternalism, however, implies a strictly minimalist self, in which there is no substantive content that can be undermined by the content of our choices. By making mental competence the sole criteria for whether a patient's

informed consent has authority as an exercise of autonomy, the approach demonstrates a 'process standard' of autonomy. That is, effective self-authorship is treated as being merely a product of the process by which the choice is reached, with no restrictions upon the actual content of that choice. Either the self has no substantive qualities to speak of, or those qualities are entirely determined by the person's operative desires, making the pursuit of autonomy merely a matter of properly informed competence and adequate opportunity.

The starting point for unpacking the concept of personal integrity is our identification with, and attribution of value to, our more central goals and values. The self, insofar as it matters to morality and autonomy, is that feature of our perspective that delineates the experiences with which we identify from those that we experience as being external impositions upon ourselves. Even if we adopt a minimalist sense of self, we can be alienated from our choices when internal or external impediments distort the choice such that it is no longer voluntary or no longer the same choice as that we are trying to make (Feinberg 1984, 116–120). As we adopt a more robust sense of self, whereby the self gains substantive traits such as attitudes and values, it becomes possible to be alienated from the *content* of a choice where it undermines these substantive traits.

I follow Wolf (1989, 141) in calling the positing of a substantial central self the 'Deep Self View', or 'DSV'. The idea of 'positing' a self needs some disclaimer. Obviously it is not a term that ought to be applied to describe our brain processes at a biological level. Nor does the DSV commit us to the claim that the deep self is a metaphysical *entity*, rather than merely a psychological concept that we use in order to make sense of ourselves and others (Frankfurt 1988, 11–25; Frankfurt 1999, 129–141). There have

been numerous philosophical explanations of how a DSV may operate, the most famous being the 'endorsement' model developed by Frankfurt (1988, 11–25) and G Dworkin (1989, 54–62). On that model, a person has higher-order desires and volitions, i.e. desires and volitions about lower-order desires, such as a desire to stop desiring alcohol. The goals most central to a person's identity are determined by one's higher-order volitions, such as a commitment to promote or combat a particular desire, or to prioritise a particular goal. Variations on the DSV are plentiful. Meyers (2005) writes that restricting autonomy to second-order endorsement wrongly excludes autonomy arising from one's embodied, relational and social self. Meyers reasons that our autonomy is sometimes enhanced in ways that don't involve deliberative endorsement, such as where one's body instinctively reacts to better pursue a physical goal, or where one's social relationships enhance one's resolve. Velleman (2005) talks of the self as a narrator, again not as an actual entity, but as (at 65) 'the reflective representation that feeds back into the person's behavior.' Oshana (2005, 78) talks of the self as one's central identity traits, i.e. 'the characteristics and relationships that are integral to a person's nature, motives, and life-plans.' Christman (2005, 345) talks instead of retrospective validation, wherein 'were one to turn a reflective eye toward the motives, values, and concepts that structure one's judgments (and do so in a piecemeal manner), one would not feel deep self-alienation, self-repudiation, and unresolvable conflict.' In the context of paternalism, Quante (1999) talks about respecting the most authentic manifestation of a person's will, and Widdershoven and Berghmans (2001, 96) emphasise the person's 'fundamental values'.

The diversity of these descriptions reflects the difficulty of expressing an intuitively simple but opaque concept as a formula. They are different descriptions of the same process, where that process is easily understood as a loose concept but nonetheless

defies precise definition. We know through experience what it is to attribute greater normative authority to one goal than another, and what it means for a value to be held only trivially, and the lack of philosophical consensus over how to put this experience into exact words should not deter us from applying it as a shared concept.

A common feature of these accounts is that they adopt the perspective of the person whose self is being examined. The boundaries of our volitional self are defined by our processes of identification and decision-making. The gate-keepers of medical paternalism – the doctors and other medical staff who assess the patients and order the intervention - are not in such a position. When determining whether a particular choice is truly autonomous, the paternalist does not have direct access to the person's will, and cannot directly observe whether the choice is only instrumentally valuable or is a deep goal valued for its own sake, or what further deep goal confers instrumental value upon the choice. To the extent that autonomy grounds moral rights against paternalism, the paternalist is at a disadvantage.

From this quandary has arisen a quite different set of criteria regarding the determination of a person's deep goals. These are the practical criteria that seek to best estimate the goals that are most central under the DSV: i.e. the person's 'settled' or 'stable' goals, 'long-held' goals, those which the person has had substantial opportunity to reflect upon, or those desires that the person holds when calm, reflective, competent, sane and not under excessive pressure (Stirrat and Gill 2005, 130; Archard 1993, 348; Kleinig 1983, 72). None of these criteria provide an exact measure of whether a desire is one of a person's deep goals. Authentic realisations of what one deeply cares about may arise unexpectedly and suddenly, and it would be absurd to suggest that there is some arbitrary minimum time delay before a person can

truly value something. Not all of our deep goals require extensive reflection or deliberation, and on some occasions it can be our deeply emotional moments that reveal to us what our most deeply held goals are. Nonetheless, such criteria provide a decent 'rough' guide. They reflect the circumstances in which we have most confidence that a person's choice is consistent with her deep goals. Sudden changes, or a lack of stability in a person's goals, don't necessarily mean that the new goals aren't authentic, but they do give us good reason to at least delve further. Similarly, one will rarely get a *less* accurate account of a person's deep goals if one asks the person to take time to calmly reflect upon her decision.

5.4 A problem of perspective

The informed consent doctrine presents autonomy as a feature of individual choosers, when it is often a feature of *relationships*. Even though mental competence is contextual, we use it to measure the mental capacities of an individual in relation to a choice. Other people can provide the information needed for informed consent, or enhance the chooser's competence by providing assistance, but competence still measures the person's mental capacities. The implication is that autonomy is achieved by adding adequate information to adequate brain processes, and removing any barriers to the proper functioning of those brain processes. On this standard view, we may need assistance in obtaining the relevant information but, once armed with that information, autonomy is something that we can achieve on our own, through the adequacy of our own mental processes. This is autonomy as it applies to the kind of decision that philosophers use to illustrate instrumental reasoning at its simplest – deliberate choices that do not require any particular expertise, that will not draw social criticism, that are not greatly influenced by our emotional reactions and that do not require any immediate self-sacrifice. That is, choices where we simply pursue an end

using the logically most efficient means: all we need for autonomy here is an authentic end, enough information and mental capacity to determine a means of pursuing that end, and freedom from anything that might get in the way.

Outside of this category, autonomy becomes a little more nebulous. Consider quitting a habit, such as smoking or biting one's nails. We might exercise autonomy over the initial commitment in the aforementioned way, but in order to impose that choice we must make a series of spontaneous decisions that are not entirely deliberate. We may need encouragement from friends, surroundings that reinforce our original choice, reminders of our motivations, and maybe even the occasional paternalistic nudge such as the friend who confiscates the packet of cigarettes bought in a moment of weakness. As Meyers (2005, 42–48) notes, autonomy here has social and physical – maybe even environmental dimensions.

People in most current societies are massively interdependent. The knowledge that makes our way of life possible – knowledge of electricity generation, engine maintenance, organisational management, agricultural techniques, software design, building construction – is divided among many people through specialisation. Each of us possesses only a fraction of this knowledge, and we are all dependent on relationships of professional and technical assistance. To make sense of our environment, where that includes our social and technological environment, we need to rely on others. In the medical context, the exercise of autonomy is a joint practice between oneself and an entire network of others who conduct research, study medical theory, conduct laboratory tests, provide medical examinations and so forth. The traditional account of mental competence takes their contribution into account as a means of obtaining information, but the medical staffs' role in making possible the

process of a patient reaching an autonomous decision goes much further: it involves the carrying out of tests, analysing of results, studying of medical theory and the *creation* (not just knowledge) of opportunities to pursue the patient's goals. The weighing up of probabilities, such as determining what risk of cancer is sufficient to make immediate testing a priority, may itself be an exercise of professional skill rather than mere 'information' to be provided to the patient. The number of marginally useful tests and statistically possible illnesses may often exceed the financial resources available, not to mention that the patient herself is likely to have limited patience and time, and for the information to be useful to a patient the doctor needs to make a clinical judgment about which of the available information is *significant*.

In this context, the traditional mental competence framework seems to draw a somewhat arbitrary distinction. Very few of us outside of the medical professions could effectively make the kind of informed choice that the framework envisages, one where the patient is provided with the information relevant to the decision and then proceeds to make an independent choice in pursuit of the way of living to which she aspires. In practice, informed consent is a much narrower decision: the medical staff recommends a treatment plan in accordance with the medical institution's policies and resources, and the patient can either accept or refuse the treatment on the terms that it is offered. The patient's consent is an exercise of trust, rather than reason. This kind of 'take it or leave it' consent is far from an adequate provision of patient autonomy, as O'Neill (2002, 43–60; 2003) has discussed at length, but even if we expand greatly the communication and options open to the patient, the patient is still ultimately reliant upon others to exercise expertise that she does not possess. The theoretical criteria for competence in medical treatment presents a standard that hardly anyone could meet, whereas the actuality involves a different process altogether.

The role of judgment and expertise in medical reasoning means that patient autonomy is, not so much a process of receiving information on various options, but rather one where the processes of reasoning are performed by medical staff on the patient's behalf. The patient is reliant upon engagement with medical professionals, and the patient's autonomy requires a treatment relationship of well-grounded trust. The medical decision-making process, as a shared project between the patient and medical staff, must itself be authentic with regard to the patient's attribution of value. But authenticity here doesn't mean mirroring the patient's own judgment – the very purpose of the treatment relationship is that the patient's judgment is augmented by the treating team's expertise. Instead, it is the principles and the process by which the decision-making process is governed that matters. Firstly, the normative principles that provide the grounds for evaluative judgment – the process's motivations, goals and values – must be those authentic to the patient. Secondly, the relationship itself must be consistent with those goals and values.

The social dimension of autonomy provides that autonomy can be meaningfully pursued, albeit imperfectly, even if we are incapable of doing so independently, and sometimes even through paternalistic measures. Even outside of contexts that require expert judgment, our attribution of value sometimes supports paternalistic intervention by those close to us in order to enable us to effectively pursue our goals and values despite weakness of will and other ordinary mental limitations. The gatekeepers of psychiatric paternalism, of course, are not friends or loved ones of their patients, and lack that informal relationship of mutual consent to the cooperative exercise of agency. In voluntary treatment, this is not a problem as the patient is willingly entering into a

formalised relationship of care that provides for a similar sharing of agency. Where treatment is imposed through the coercion of psychiatric detention or involuntary treatment, the patient actively rejects the authenticity of the treatment relationship, and without a pre-existing relationship of trust, the analogy to intervention by close friends breaks down.

Nonetheless, there are a great many of us who hope that in the event that we were to risk catastrophic self-harm due to mental illness, medical institutions would paternalistically intervene if necessary. Moreover, we hold goals and values that, to us, are more important than the preservation of sovereignty over our medical treatment and, if the threat to those goals is sufficiently severe, perhaps even our short-term freedom. The relevance of the friendship relationship that makes it morally permissible for a friend to confiscate a packet of cigarettes, is not that we provide some implied consent to this paternalism (we often don't) or that we will subsequently consent (even if it is possible to consent 'after the event', this would leave the morality of intervention ambiguous at the time of action), but that it is a practice embedded in our cultural environment, such that it provides part of the social machinery that makes meaningful autonomy possible.

Paternalistic responses to self-harm arising from mental illness have a surprisingly broad cross-societal existence. Horwitz's comprehensive review (1982, 92) notes that paternalistic practices such as locking the person inside her home (albeit for shorter periods, and often with less social stigma, than common to Western societies) have been identified in sociological studies of indigenous peoples from Mexico, the Arctic, east Africa, Ghana, New Guinea, Sri Lanka and the Caroline Islands. Reznek (1991, p.126) notes that there is some evidence that in the 14th and 15th centuries, the feudal

system in Western Europe provided a similar system of care, whereby the seriously mentally ill would be transported back to their own villages to be cared for by the combined village (while citing evidence from Maher and Maher (1982) that Foucault's (2001) memorable description of 'ships of fools' crewed by enslaved teams of the mentally ill, supposedly prior to the invention of psychiatric institutions, lacks even marginal historical evidence). The cross-cultural reach of psychiatric paternalism, and its informal practice by family and friends in the absence of formal state institutions, reflects that mental illness and consequent loss of the capacity for autonomy, are part of the basic challenges of social existence that virtually any functional society must address.

In western society, giving the paternalistic function, and its associated authority, to family or friends brings its own potential for abuse. These relationships can themselves include substantial differences in values, and where the intervention moves beyond the minor and relatively harmless, it can become a tool for the wrongful imposition of one person's values upon another. Expectations of shared value systems across families can exacerbate this, as parties may be tempted to use the paternalistic authority without bona fides, our may be more inclined to the formation of different goals and values by members of the family as the result of illness. In a pluralistic society, we are committed to preventing this imposition of goals and values alien to the patient. Moreover, the means of psychiatric intervention (even the short-term detention found in many indigenous cultures) is far from benign, relying upon the safety and efficacy of pharmaceutical products, and the force of the state (or village) to physically detain and coerce the patient. It is difficult to see how a liberal society could permit intervention of this kind, except through a system that is heavily supervised by the state. The state is, in effect, supplanting the pre-existing cultural practices in an

endeavour to both provide better treatment, and to protect the patient from imposition of others' goals and values. Given that mental illness is a normative classification, but also a real and sometimes necessary one, societal intervention is part of the social mechanisms that allow us to pursue a way of living that is authentic to our goals and values. The function of liberal ethics is not to prevent such intervention, but to ensure that the evaluative judgments that it involves allow for a meaningful pluralism of self-oriented goals and values.

With the patient's goals and values providing the guiding principles for intervention, it is irrelevant whether a patient whose values are at odds with state intervention is immediately capable of making an effective treatment choice. We respect her autonomy by allowing her to sacrifice her well-being for her freedom from intervention, as per her attribution of value. Similarly, we respect the autonomy of a patient whose refusal of treatment arises from her delusional beliefs, rather than her goals and values, *by* intervening contrary to her immediate decision. This begs a theory of personal identity that allows for a moral distinction between a person's immediate values and her authentic identity. Part of this is addressed by the theory of deep self, in which a person's reflective and higher-order values are ordinarily taken to be more deeply representative of her identity. However, we also require an account of how change to one's values affects authenticity and identity. I endeavour to provide that account in ‘Beyond Mental Competence’ (Edwards 2010).

It is not enough that the institutions of psychiatric paternalism have the right goals (i.e. to protect the patient's own goals, values and desired way of living). The style and methods of the treatment relationship must also be authentic with regard to these goals. The paternalist must have the facilities and other resources that are needed in order to

provide treatment with dignity, and for the patient to live the way of life that is authentic to her values. This extends from the general comfort of treatment facilities to the provision of specific religious and cultural services, and is a necessary part of the paternalist's responsibility, rather than an obligation to pursued 'as much as possible' after the decision to intervene has already been made. Nonetheless, there is a comparative dimension to this – treatment facilities may be inadequate, but comparatively far more dignified than the patient's current accommodation if she is effectively being coerced into homelessness due to the severity of her illness. However, unless a substantial effort is made by the state to meet these needs, its bona fides in endeavouring to respect the patients' autonomy is suspect. Similarly, it is difficult to see how an institution can be meaningfully protecting a patient's goals and values if hospitalisation becomes a means for 'holding' patients under conditions where they are free from physical self-harm, but unable to live a life to which they can attribute meaning. There may well be a minimum required standard of care for psychiatric paternalism that is far higher than simply providing a net benefit to the patient's well-being.

If the works of Feinberg (1984, 68) and Arneson (1980) are any indication, some people do not want paternalism to play *any* role in their medical relationships. The flexibility of the appeal to personal integrity is such that it can accommodate those for whom the authentic life is one of personal sovereignty, where the patient is free to make her own mistakes, no matter how disastrous to her other goals and values. If the demand for personal sovereignty genuinely reflects the patient's attribution of value, then freedom itself must be her most central value by so great an extent that she would sacrifice all of her other goals and values rather than impose upon it. Most of us, I suspect, do not value freedom so highly. We need some freedom in order to achieve

our other central goals and values, and we may value such independence for its own sake, but our own values would still suggest that there are times that our freedom should be sacrificed to preserve our more important pursuits.

Finally, I acknowledge that some of the most severely impaired patients are either unable to form goals and values, are so incapable of self-awareness that they are unable to identify with those goals and values, or are subject to such fluctuations in attitude that they lack a persistent character. This necessitates a different, capacity-related, form of paternalism. Patients who are mentally incompetent in this sense, where they are literally unable to form the 'self' that is at the centre of personhood, simply lack any moral interest in autonomy (although they may benefit from some of the consequential interests that stem from freedom).

5.5 The appeal to personal integrity in practice

When people engage with large medical institutions, they do so from an exceptionally vulnerable position and on coerced terms. The engagement is governed almost entirely by the institution's own policies: the individual patient, having been forced by illness to engage with the institution, must comply with the institution's mechanisms for referral and provision of treatment, as determined by either the bureaucracy of the state or the institution's financial interests (Bluhm 2009, 141–142). The patient is at a disadvantage in terms of information (both about her condition and the workings of the institution itself), financial means and, legal authority, in addition to the mental and physical constraints of their illness. This imbalance is at its greatest for hospital patients. Hospitals are controlled environments, where one can only effectively pursue those goals and values that are consistent with the programmes and policies of the institution itself. Instead of the engagement being merely an interruption to, or a

restriction upon, the patient's way of life, hospitalisation replaces that way of life, wholly integrating the patient into the institution's workings. Through this, medical institutions provide a mechanism, with a reach that is rare in liberal societies, for broad-ranging imposition by the state upon aspects of peoples' lives that are usually considered personal. Ordinarily, the state does not concern itself with whether a person is religiously opposed to blood transfusions, or whether an individual believes that electro-convulsive therapy ('ECT') or antipsychotic medication does more harm than good. Moreover, the state would be hard-pressed to actually enforce such regulation, without resorting to drastically authoritarian measures. However, when that person is forced into hospital by the severity of her illness, the state can determine whether the person is able to live in accordance with her own beliefs and values, without that person having any practical means of avoidance.

Patient autonomy requires relationships of care in the context of a medical institution that makes them exceptionally vulnerable to authoritarian imposition. We need a system of protections against paternalism that strengthen the authority of the patients' goals and values, without pretending that the patient can achieve autonomy by disassociating herself from her networks of medical care and advice. The task is made more difficult by the opaqueness of character. Our knowledge of another's goals and values is imperfect, and that is especially so when we only encounter a person during a time of severe mental impairment. Very often, medical staff will be in no position to declare that the patient's central goals and values differ from her operative desires: if the full extent of their knowledge of a patient is taken from her immediate presentation and information from family members (who may themselves be unaware of the patient's central values, or may even deliberately seek to override those values), then a reliable estimate of the patient's deep values is unlikely. The competence approach

may provide a pragmatic compromise, working from the premise that *usually* a competent patients is best placed to know her own interests.

However, a great deal of medical treatment does not match this state of affairs. Most treatment for both mental and physical illness is provided through general practitioners, who are well-placed to encourage a long-term treating relationship. Psychiatric treatment often extends past the initial hospitalisation through outpatient services or inpatient residential care, carried out over a time period that may run for many years. Even if the competence approach cannot practically be replaced from the ground up, we have no reason not to allow doctors to apply the greater knowledge and trust from a long-term treating relationship into a more precise paternalistic tool. In particular, we ought to guarantee that patients are never denied respect for their personhood on the basis that their decision seems grossly unreasonable to medical staff or society in general. If autonomy turns on relationships conducive to the patient's values and goals, then the reasonableness of their informed consent or refusal of treatment ought be judged in light of those values and goals, whenever such an assessment is possible. This in itself should address the practice of conflating content and competence, whereby minority cultures and eccentric individuals are denied personhood on the basis of their decisions rather than their capacities.

It also suggests a more sensible approach to psychiatric paternalism in the difficult mood disorder cases. Provision of emergency treatment to a previously unknown patient is one thing, but longer term paternalism, as may be imposed post-discharge from involuntary hospitalisation through Community Based Orders, ought consider the patient's character, and the way of life she wants to live, before determining that her condition makes it impossible for her to autonomously choose (or refuse) treatment.

Futhermore, at a time when advance directives are encouraged and used to refuse *life-saving* treatment, we have no excuse for ignoring a patient's long-held views on psychiatric treatment the moment she becomes mentally incompetent.

In the introduction to this section, I described as an absurdity the practice of delaying paternalistic intervention during a patient's predictable but preventable deterioration, not because the patient holds any ideological, cultural or personal value that prohibits psychiatric treatment, but because the patient's loss of autonomy is not immediately attributable to mental incompetence. Mental competence is misleading in such cases, because the capacity for autonomy is reliant upon engaging in a treatment relationship that makes autonomous decision-making possible. Few, if any, of us are ever 'competent' to choose or refuse treatment in the absence of such a relationship, and when threatened with self-harm so drastic that it outweighs the abundant benefits of free choice, it is disingenuous to turn to autonomy as a ground for refraining from intervention.

Relationships are not easily shaped through formal codes of rights and responsibilities. Such codes deal in benchmarks and enforceable requirements, whereby compliance requires only the practitioners' diligence and professional competence. To be of use, rights must be capable of measurable enforcement – a requirement that favours the protection of negative liberties over the fostering of positive ones. It is difficult to see how a formal ethical code, or a series of legal rights, can enforce a demand for supportive treatment relationships. The appeal to personal integrity is better served by making the protection of personal integrity one of the values upon which involuntary treatment policies are designed, supporting such piecemeal protections as acknowledging the authority of properly made advance directives, and rights to the

judicial or quasi-judicial review of all paternalistic intervention.

The positive requirements for personal integrity are much greater than that which can be practicably imposed through legal rights limiting paternalism. Limitations upon paternalism do not give practitioners the time to develop a better knowledge of their patients, nor does it guarantee concern for a patient's values, nor the provision of treatment options that patients of diverse values and cultures can identify with. Nonetheless, we do have a precedent for the partial protection of personal integrity through negative liberties in the implementation of anti-discrimination and human rights legislation. Like the appeal to personal integrity, prevention of unjust discrimination is not conducive to narrowly defined rights. It relies on discretionary authority and the application of judgment as to what discrimination is unjust, firstly by those whose decisions are subject to the legislation, and then upon quasi-judicial and judicial review, and in the executive branch's prosecution function. The current protections arising from mental competence could be more justly replaced with a simpler, broader rule, added to the existing requirements regarding the severity of the relevant self-harm: *that involuntary treatment must not violate a patient's cultural or personal values*.

The centrality of mental competence to paternalism is not solely a legislative provision, but it underlies our system of involuntary treatment, defining the aims that we seek to achieve through our policy on medical paternalism. It is a flawed approach, purporting to protect a freedom of personal sovereignty, when most of us live in a social environment where autonomy can only be effectively pursued through interpersonal relationships. This flaw has led to practical injustice, such as the inability to intervene early in 'revolving door' relapses, and the lack of legal authority attributed

to psychiatric advance directives. The protection of autonomy and personal integrity lies in fostering the right kind of treatment relationships, governed by the patient's deep-held goals and values, and this ought to provide the primary aim that underlies the criteria for involuntary treatment.

References

Archard, David. 1998. 'Informed Consent: Autonomy and Self-Ownership'. *Journal of Applied Philosophy* 25 (1): 19-34.

———. 1993. 'Self-Justifying Paternalism'. *Journal of Value Inquiry* 27: 341-352.

Arneson, Richard. 1980. 'Mill Versus Paternalism'. *Ethics* 90 (4): 470-489.

Beauchamp, Tom, and James Childress. 2009. *Principles of Biomedical Ethics*. 6th ed. New York: Oxford University Press.

Bluhm, Robyn. 2009. 'Evidence-based Medicine and Patient Autonomy'. *International Journal of Feminist Approaches to Bioethics* 2 (2): 134-151.

Boddington, Paula. 1998. 'Organ Donation After Death - Should I Decide, or Should My Family?' *Journal of Applied Philosophy* 13 (1): 69-81.

Botha, Ulla, Liezl Koen, John Joska, John Parker, Neil Horn, Linda Hering, and Piet Oosthuizen. 2010. 'The Revolving Door Phenomenon in Psychiatry: Comparing Low-frequency and High-frequency Users of Psychiatric Inpatient Services in a Developing Country.' *Social Psychiatry and Psychiatric Epidemiology* 45 (4): 461-468.

Brock, Dan. 1988. 'Paternalism and Autonomy'. *Ethics* 98 (3): 550-565.

Buchanan, Allen, and Dan Brock. 1986. 'Deciding for Others'. *The Millbank Quarterly* 64 (Supplement 2: Medical Decision Making for the Demented and Dying): 17-94.

———. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*. Cambridge: Cambridge University Press.

Cale, Gita. 1999. 'Continuing the Debate Over Risk-Related Standards of Competence'. *Bioethics* 13 (2): 131-148.

Charland, Louis. 2002. 'Cynthia's Dilemma: Consenting to Heroin Prescription'. *The American Journal of Bioethics* 2 (2): 37-47.

Checkland, David. 2001. 'On Risk and Decisional Capacity'. *Journal of Medicine and Philosophy* 26 (1): 35 – 59.

Christman, John. 2005. 'Autonomy, Self-Knowledge and Liberal Legitimacy'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 330-358. Cambridge: Cambridge University Press.

Culver, Charles, and Bernard Gert. 1990. 'The Inadequacy of Incompetence'. *The Millbank Quarterly* 68 (4): 619-643.

DeMarco, Joseph. 2002. 'Competence and Paternalism'. *Bioethics* 16 (3): 231-245.

Dworkin, G. 1989. 'The Concept of Autonomy'. In *The Inner Citadel: Essays on Individual Autonomy*, ed. Christman. New York: Oxford University Press.

Dworkin, Gerald. 1971. 'Paternalism'. In *Morality and the Law*, ed. Richard Wasserstrom, 107-126. Belmont: University of Minnesota Press.

Dworkin, Ronald. 1986. 'Autonomy and the Demented Self'. *The Millbank Quarterly* 64 (Supplement 2): 4-16.

———. 1993. *Life's Dominion*. New York: Knopf.

Edwards, Craig. 2009. 'Ethical Decisions in the Classification of Mental Conditions as Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 73-90.

———. 2010. 'Beyond Mental Competence'. *Journal of Applied Philosophy* 27 (3): 273-289. Faden, Ruth, and Tom Beauchamp. 1986. *A History and Theory of Informed Consent*. New York: Oxford University Press.

Feinberg, Joel. 1984. *Harm to Self*. Vol. 3. 3 vols. The Moral Limits of the Criminal Law. New York: Oxford University Press.

Foucault, Michel. 2001. *Madness and Civilisation: A History of Insanity in the Age of Reason*. Trans. Richard Howard. London: Routledge.

Frankfurt, Harry. 1988. *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press.

———. 1999. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Friedman, Marilyn. 2005. 'Autonomy and Male Dominance'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 150-173. Cambridge: Cambridge University Press.

Grisso, Thomas, and Paul Appelbaum. 1995. 'The MacArthur Treatment Competence Study'. *Law and Human Behaviour* 19 (2): 105-126.

Hill, Thomas. 1989. 'The Kantian Conception of Autonomy'. In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, 91-105. New York: Oxford University Press.

Horwitz, Allan. 1982. *The Social Control of Mental Illness*. New York: Academic Press.

Kastrup, Marianne. 1987. 'Who Became Revolving Door Patients?' *Acta Psychiatrica Scandinavica* 76: 80-88.

Kemshall, Hazel, Louise Marsland, Thilo Boeck, and Leigh Dunkerton. 2006. 'Young People, Pathways and Crime: Beyond Risk Factors'. *The Australian and New Zealand Journal of Criminology* 39 (3): 354-370.

Kihlbom. 2008. 'Autonomy and Negatively Informed Consent'. *Journal of Medical Ethics* 34: 146-149.

Kirby, Michael. 1983. 'Informed Consent: What Does It Mean?' *Journal of Medical Ethics* 9: 69-75.

Kleinig, John. 1983. *Paternalism*. Totowa: Rowan and Allenheld.

Kluge, Eike-Henner. 2005. 'Competence, Capacity and Informed Consent: Beyond the Cognitive-competence Model''. *Canadian Journal on Aging* 24 (3): 295-304.

Kress, Ken. 2006. 'Rotting with Their Rights On: Why the Criteria for Ending Commitment or Restraint of Liberty Need Not Be the Same as the Criteria for Initiating Commitment or Restraint of Liberty, and How the Restraint May Sometimes

Justifiably Continue After Its Prerequisites Are No Longer Satisfied'. *Behavioral Sciences and the Law* 24: 573-598.

MacDonald, Robert. 2006. 'Social Exclusion, Youth Transitions and Criminal Careers: Five Critical Reflections on "Risk"'. *Australian and New Zealand Journal of Criminology* 39 (3): 371-383.

MacIntyre, Alasdair. 1981. *After Virtue*. London: Duckworth.

Maclean, Alasdair. 2000. 'Now You See It, Now You Don't: Consent and the Legal Protection of Autonomy'. *Journal of Applied Philosophy* 17 (3): 277-288.

Maher, Winnifred, and Brendan Maher. 1982. 'The Ship of Fools: Stultifera Navis or Ignus Fatuus'. *American Psychologist* 37: 756-761.

Matthews. 2000. 'Autonomy and the Psychiatric Patient'. *Journal of Applied Philosophy* 17 (1): 59-70.

Mental Health Act 1996 (WA)

Mental Health draft Bill for community discussion 2011 (WA) Government of Western Australia Health Commission, Health Commission Homepage: *http://www.mentalhealth.wa.gov.au/mentalhealth_changes/mh_legislation.aspx*, last modified 14 March 2012, checked 14 March 2012.

Meyers, Diana. 2005. 'Five Faces of Selfhood'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 27-55. Cambridge: Cambridge University Press.

O'Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

———. 2003. 'Some Limits of Informed Consent'. *Journal of Medical Ethics* 29: 4-7.

Oshana, Marina. 2005. 'Autonomy and Self-Identity'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 77-97. Cambridge: Cambridge University Press.

Quante, Michael. 1999. 'Precedent Autonomy and Personal Identity'. *Kennedy Institute of Ethics Journal* 9 (4): 365-381.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Rigakos, George. 1999. 'Risk Society and Actuarial Criminology: Prospects for a Critical Discourse'. *Canadian Journal of Criminology* 41 (2): 137-150.

Rogers v Okin 478 F.Supp. 1342 (D. Mass. 1979), 634

Rudnick, Abraham. 2002. 'Depression and Competence to Refuse Psychiatric Treatment'. *Journal of Medical Ethics* 28: 151-155.

Sandman, Lars. 2004. 'On the Autonomy Turf. Assessing the Value of Autonomy to Patients''. *Medicine, Health Care and Philosophy* 7: 261-268.

Scoccia, Danny. 1990. 'Paternalism and Respect for Autonomy'. *Ethics* 100 (2): 318-334.

Sell v United States, 539 US 166 (2003)

Spellecy, Ryan. 2003. 'Reviving Ulysses Contracts'. *Kennedy Institute of Ethics Journal* 13 (4): 373-392.

Stanley, Barbara, and Michael Stanley. 1982. 'What Is It? How Is It Assessed? Testing Competency in Psychiatric Patients''. *IRB: Ethics and Human Research* 4 (8): 1-6.

Stirrat, Gordon, and Robin Gill. 2005. 'Autonomy in Medical Ethics After O'Neill'. *Journal of Medical Ethics* 31 (3) (March): 127-130. doi:10.1136/jme.2004.008292.

Taylor, Charles. 1989. *Sources of the Self: The Making of the Modern Identity*. Cambridge: Harvard University Press.

Taylor, and Hawley. 2006. 'Health Promotion and the Freedom of the Individual'. *Health Care Anal* 14: 15-24.

VanDeVeer, Donald. 1980. 'Autonomy Respecting Paternalism'. *Social Theory and Practice* 6 (2): 187-207.

———. 1986. *Paternalistic Intervention: The Moral Bounds on Benevolence*. New Jersey: Princeton University Press.

Velleman, David. 2005. 'The Self as Narrator'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 56-76. Cambridge: Cambridge University Press.

Wicclair, Mark. 1999. 'The Continuing Debate Over Risk-Related Standards of Competence'. *Bioethics* 13 (2): 149-153.

Widdershoven, Guy, and Ron Berghmans. 2001. 'Advance Directives in Psychiatric Care: a Narrative Approach'. *Journal of Medical Ethics* 27: 92-97.

Wikler, Daniel. 1983. 'Paternalism and the Mildly Retarded'. In *Paternalism*, 83-115. Minneanapolis: University of Minnesota Press.

Wilks, Ian. 1997. 'The Debate Over Risk-Related Standards of Competence'. *Bioethics* 11: 413-426.

———. 1999. 'Asymmetrical Competence'. *Bioethics* 13: 154-159.

Wolf, Susan. 1989. 'Sanity and the Metaphysics of Responsibility'. In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, 137-151. New York: Oxford University Press.

Part B

Paper 4 - Suicide prevention and the limits of patient autonomy

Abstract: In the previous two essays, I have analysed the case for moral rights against psychiatric paternalism, that stand irrespective of whether intervention would benefit the patient. I began with the liberal traditions that underlie the modern concept of patient autonomy, and proceeded to argue for a revision of the way this concept is encapsulated in the applied context of bioethics. In criticising the practical liberal doctrines of mental competence and informed consent, I nonetheless have put an account that remains thoroughly liberal in that it takes moral authority from the patient's own attribution of value. In this paper, I complete this part of my project by examining the limits of patient autonomy. I inquire into the scope of the rights that can be derived from a liberal conception of patient autonomy, and whether an unadulterated liberal respect for patient autonomy necessarily commits us to refrain from paternalistic interference in the patient's informed and mentally competent choices. In this manner, I am testing the ultimate implications of the account of liberal restrictions upon autonomy that I have been developing in the past three essays. Last essay, that development culminated in my claim that mental competence, as a necessarily evaluative concept, is of less importance to medical decision-making than the authentic pursuit of a patient's values. In this paper, I argue that once we recognise that mental competence cannot be normatively neutral, we cannot maintain the simplistic exclusion of hard paternalism that has come to define much of liberalism on paternalism. Instead, we must put forward a positive account of personhood to meet authoritarian claims directly, by addressing the shared concern of human function or flourishing. In the process, I develop a defence of hard paternalism in relation to suicide (or more accurately, one form of suicide), and illustrate how the more robust

account of personhood developed last essay demands harder limits upon the boundaries of personal autonomy.

Suicide prevention and the limits of patient autonomy

1. Psychiatric paternalism in the prevention of suicide: an exception to legal and ethical principles.

In the previous two essays, I have analysed the case for moral rights against psychiatric paternalism, that stand irrespective of whether intervention would benefit the patient. I began with the liberal traditions that underlie the modern concept of patient autonomy, and proceeded to argue for a revision of the way this concept is encapsulated in the applied context of bioethics. In criticising the practical liberal doctrines of mental competence and informed consent, I nonetheless have put an account that remains thoroughly liberal in that it takes moral authority from the patient's own attribution of value. In this paper, I complete this part of my project by examining the limits of patient autonomy. I inquire into the scope of the rights that can be derived from a liberal conception of patient autonomy, and whether respect for a patient autonomy necessarily commits us to refrain from paternalistic interference in the patient's informed and mentally competent choices.

In Western societies, the traditional moral condemnation of suicide has progressed into a institutional willingness, by all appearances supported by widespread social sentiment, to view the prevention of suicide as a proper target of public policy. In the medical context, this has manifested in a broad program of involuntary psychiatric treatment that is aimed at the prevention of suicidality. I suspect that for most people, this involuntary treatment of suicidal intent (insofar as it concerns those who are physically healthy) is utterly unremarkable, save for the desire that greater funding and effort were allocated towards it. Nonetheless, from the perspective of liberal theory on

medical paternalism, the practice – and the social support it enjoys – is quite remarkable. It operates within medical institutions that are ethically and legally governed by the requirement that all treatment be contingent upon the patient's informed consent, unless the patient is mentally incapable of consent and refusal (Kirby 1983; Kihlbom 2008). Yet the practice itself, and the theory upon which it operates, appears to exclude entirely the possibility of an informed and mentally competent refusal to partake in treatment. The very fact that a patient is at risk of suicide is viewed as reason enough to apply psychiatric treatment, paternalistically if necessary.

While the authoritative Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association 2000) does not list 'suicidal intent' as an illness in itself, it forms part of the diagnostic criteria for both minor (American Psychiatric Association 2000, chap. Appendix B: Minor Depressive Disorder) and major depressive disorder (American Psychiatric Association 2000, chap. Mood Disorders). More importantly, while suicidal intent is not formally sufficient for mental illness nor mental incompetence, texts on psychiatric practice strongly advocate in-hospital detention as a precautionary measure, for all potential suicidality, not just those secondary to another mental illness, for example:

> *'Ensuring safety is the first and most important step in managing a potentially suicidal patient. This can best be accomplished through admission to a secure patient holding area. If no such facility is available, close observation of the patient by a staff member is another option. In this circumstance, moving the patient to a more secure location is the first treatment goal.' (Weigel et al. 2009)*

The American Psychiatric Publishing Textbook of Psychiatry is similarly assertive of the obligation to ensure suicidal patients receive immediate psychiatric treatment, through involuntary hospitalisation if necessary:

> '*If the patient rejects the clinician's recommendation for hospitalization, the matter is immediately addressed as a treatment issue. Because the need for hospitalization is acute, a prolonged inquiry into the patient's reasons for rejecting the recommendation for hospitalization is not feasible. Furthermore, the therapeutic alliance may be strained. Consultation and referral are options for the clinician to consider, if time and the patient's condition permit...The failure to involuntarily hospitalize a patient judged to be at high risk for suicide who subsequently attempts or completes suicide may result in a malpractice suit against the clinician.'* (Simon 2008).

On the face of things, this may appear to be the natural conclusion of the legislative demand to intervene where mental illness will lead to a patient's physical harm (e.g. s26 *Mental Health Act* (WA) 1996, s3 *Mental Health Act* (UK) 1983, s349.467 *Florida Mental Health Act* 1971). However, as I explained in the 'Mental competence and its limitations', the role of mental competence envisaged in psychiatric texts is build into the legislative accounts of mental illness, thus calling upon the concept of mental competence to separate mental illnesses that may well justify treatment but are innocuous for the purpose of involuntary treatment, from those that warrant stripping a person of her ability to voluntarily accept self-harm, and bringing mental health legislation at least approximately in line with the bioethics upon which it is founded (though perhaps with allowance for compromise with populism over pure ethical

theory). Moreover, this is intervention triggered *directly* by harm, rather than a risk of harm following from identification of some mental illness that exists independently of the threatened harm itself. The direct leap from suicidality to intervention is a phenomenon separate to that of the ordinary role of self-harm under civil commitment legislation.

The same assumption that suicidality directly warrants treatment, including involuntary hospitalisation, is found in the Manual of Clinical Psychopharmacology (Shatzberg, Cole, and DeBattista 2010, Depression and Suicidality). No distinction is raised regarding the patient's mental competence or any required nexus between suicidal intent and diagnosed mental illness. Nor does there appear to be any consideration of the possibility that a suicide attempt might be a mentally competent choice worthy of the protection of the informed consent doctrine. At least insofar as the influential American psychiatric community is concerned, the position is clear: if the patient refuses assistance as a voluntary patient, suicidality justifies paternalistic intervention.

Most contemporary liberal philosophers would argue that popular moral thinking on suicide is wrong in a number of ways, two common claims being that at least some of the moral distinctions we draw between active and passive euthanasia are unjustified, and that there are circumstances in which suicide can be justified in terms of the person's prudential interests (Brock 1992, 12–13). In this paper, I am not concerned with inquiring deeply into this debate about the kind of suffering or terminal illness that could justify a decision to end one's own life. Instead, my interest is to ask why the private tragedy of suicide is so widely accepted as justifying the state's paternalistic psychiatric treatment of suicidality. As it stands, psychiatric policy on suicide-

prevention gives the appearance of a rebuke to the liberalism that is normally dominant within bioethics. Despite this, there has been no great social, or even philosophical, movement in opposition to paternalistic suicide-prevention programs, except concerning cases at the very edge of the practice – suicide amongst the incurably ill or incapacitated (Cantor 1993; DeGrazia 2005, 137–170; Dworkin 1993, 228–231; Harvey 2006; Jansen 2006; Savulescu 1997; Kadish 1992; Buchanan and Brock 1989). Although the argument against paternalistic intervention in these cases is put in terms of respect for personal autonomy, the terrible pain and advancing dementia faced by these patients appear to invite an alternative explanation in terms of the patients' prudential interests, or at least in the social endorsement of the patient's condition as one which could reasonably justify a desire to die. There is a conspicuous absence of works defending suicide as an expression of autonomy in circumstances that most people would find ridiculous or tragic. If the case against paternalistic suicide-prevention is based in autonomy, rather than prudential interests, it is reasonable to ask why there is such an absence of liberals in bioethics willing to defend suicide amongst the young, healthy and opportunity-rich. After all, there are other forms of shockingly tragic self-harm, such as the final decrepit stages of heroin and alcohol abuse, where a patient's informed and competent refusal is all that is required to exclude her from any imposition of unwanted treatment. In this regard, compared to other life-threatening illnesses that may be the target of medical treatment, suicide is different.

A common response to suicide-prevention within bioethics is to try to 'have it both ways', allowing a broad program of paternalistic suicide-prevention while asserting that the prevention of suicide ought be subject to the same limits upon paternalism that are present in other parts of medical practice. Typically this entails claiming that *some* of the current medical prevention of suicide is unjust, whilst the great majority of

suicidal patients are somehow lacking the capacity for personal autonomy for reasons such as mental illness or 'inauthenticity' (where a person's judgement is inconsistent with her genuine deeply-held and self-defining goals and values, e.g. through impulsiveness, compulsion, weakness of will or confusion). Savulescu, in his short piece 'The trouble with do-gooders: the example of suicide'(1997), provides a good illustration of this view:

> *'How many people who attempt suicide competently and voluntarily do so? I have no idea. The majority of patients who attempted suicide are in good physical health. The majority have some kind of psychiatric disturbance, The level of competence required to make such an important decision ought to be set fairly high. Given this requirement, perhaps very few patients competently deserve to die. (However, this has not been systematically studied.) But this should not stop us allowing those few, like Bouvia and Mrs N, who are competent, to die if they so desire.'* (Savulescu 1997, 112)

Savulescu combines a robust defence of the application of the informed consent doctrine to suicide, with examples that seem curiously well suited to an alternative defence on prudential grounds (e.g. criticising the force-feeding of seriously and incurably ill and incapacitated patients). The same pattern appears in the works of the many bioethicists who defend or seek to extend the right of competent patients to refuse life-extending treatment (DeGrazia 2005, 137–170; Dworkin 1993, 228–231; Dworkin et al. 2007; Kadish 1992; McMahan 2002). Such examples may serve Savulescu's purpose of demonstrating the injustice of current suicide-prevention measures, but they fall far short of the full implications of his account. If the exercise of meaningful autonomy (rather than tragedy and benevolent concern) is what makes it

wrong to apply unwanted medical treatment in preventing a person's decision to die, then we must extend our concern to the liberty of people whose motivations for suicide do not elicit such sympathy.

The lack of examples involving the young, healthy and opportunity-rich may charitably be taken as an oversight. But their omission obfuscates the limitations of Savulescu's criticism of suicide-prevention. Savulescu (1997, 112) concedes that suicide-prevention is justified where the person's choice is the product of competence-denying mental illness, yet the policy of broad psychiatric intervention draws from the perception that suicidality – or, more precisely, the denial of self-worth that often motivates suicidality – is itself a manifestation of serious mental illness. Of course, this concerns only one class of suicide. Sometimes a decision to die can be a means of promoting or preserving one's deep-held goals and values, thereby affirming one's positive self-worth (Dworkin 1993, 201–206, 230; Ott 2009, 41; DeGrazia 2005, 79–84; Kuczewski 1999; Kuczewski 1994; Dworkin et al. 2007, 492). However, suicide is more commonly associated with the *denial* of self-worth, and the negation of any potential for meaning or value that could justify one's continued existence.

In this manner, the psychiatric policy of suicide-prevention addresses a concern that I suspect is widely shared in society. There is something deeply troubling about accepting the denial of self-worth as a legitimate object of personal autonomy, at least by the young, healthy and opportunity-rich. In common moral discussion, to inquire into the mental capacity or authenticity of a young and healthy, but suicidal, person's declaration of self-loathing is often to miss the point: our obligation to intervene is not alleviated by the knowledge that she is truly and authentically committed to her own worthlessness. Our concern is not that her denial of self-worth is a *consequence* of

mental impairment, but that her authentic denial of self-worth *comprises* serious mental impairment in its own right, and of a kind that renders her mentally incompetent with regard to self-harm. The DSM-IV-TR reveals that this concern (or one very much like it) is embodied within the psychiatric concept of depression. Suicidality combined with a feeling of worthlessness meets the minimum criteria for a depressive disorder (unless the symptoms are better explained by other mental illness or physiological condition, or loss of a loved one) (American Psychiatric Association 2000, Appendix B: Minor Depressive Disorder). The more that this denial of self-worth is an authentic expression of the patient's true self-conception, the more likely it becomes that she will meet the criteria for *major* depressive disorder[2] (American Psychiatric Association 2000, chap. Mood Disorders).

This illustrates the inadequacy of seeking to justify or criticise psychiatric paternalism in purely procedural terms, as though we would allow a young and healthy person to kill herself uninterrupted if only she had adequate mental capacities and clear enough thinking when she determined that her life is worthless. The test is a fraud: *nobody* within this group can ever be qualified to give informed consent, because we take their suicidality itself as proof of their mental incompetence. By taking a particular attitude, here the denial of self-worth, to be a form of mental illness, we attribute the patient with a false consciousness. Even though she identifies with this attitude, we do not acknowledge it as a part of her identity and so we invalidate the attitude as a source of reasons for action. Moreover, having identified depressive disorders as a form of mental illness, they inform our understanding of mental incompetence, such that we do

2 Five out of a possible nine criteria must be present. In abbreviated form, these are: (1) near-constant depression, (2) marked loss of interest or pleasure in one's activities, (3) unusual weight loss or gain, (4) ongoing insomnia or hypersomnia, (5) psychomotor agitation or retardation, (6) near-constant fatigue, (7) feelings of worthlessness or inappropriate guilt, (8) diminished concentration and (9) suicidality. (American Psychiatric Association 2000, chap. Mood Disorders)

not hold the person to be morally responsible or autonomous in relation to the symptoms of illness (as discussed in 'Reasons, autonomy and paternalism').

Mental illness and mental competence are both evaluative concepts whereby we impose our conception of personhood (as discussed in earlier papers in this dissertation, 'Ethical concerns in the classification of mental conditions as mental illness' and 'Mental competence and its limitations'), and we cannot hide behind the listings of the DSM-IV-TR in order to justify excluding the denial of self-worth from that conception of personhood. We need to ask the reverse of this question: whether the standard of mental competence that is implied by our suicide-prevention policies is itself consistent with a liberal conception of personhood and society.

In the rest of this paper, I explore our justifications for adopting a policy of 'hard' paternalism, i.e. intervention regardless of voluntariness and autonomy, in the prevention of suicide. The reluctance of liberal philosophers to consider the young and opportunity-rich as examples when defending the right to choose death (whether in terms of suicide or just refusal of life-saving treatment) reflects a broader division in our moral convictions on suicide. *Sometimes* the choice to die seems to justify paternalistic intervention irrespective of the patient's authenticity and lack of mental impairment, while on other occasions the choice to die appears to warrant the same liberal respect for personal autonomy as applies to other forms of self-harm. In particular, discussion of a patient's right to choose death is often framed in terms of an interest in 'dying with *dignity*'. My first major claim is that the decision to die can sometimes have a self-affirming quality, and that this affirmation of self-worth comprises at least a considerable portion of what is meant by dignity.

Secondly, I argue that we are justified in demanding that self-worth be a necessary criterion for personal autonomy. In this, I follow from my prior account of mental competence as an evaluative concept (from 'Mental competence and its limitations'). Liberal personhood is not a blank slate upon which we can impose any self-conception we have, so long as we have the right mental capacities. Rather, it affirms our capacity to imbue our own lives with meaning and value, from which we obtain our own reasons for acting. To deny one's self-worth is to deny not only the worth of one's current goals and values, but also one's capacity to seek and devise new goals and values for oneself. This capacity for reinvention and personal development is part of what makes us functional under a liberal conception of personhood.

Finally, in exploring the issue of suicide prevention, I also illustrate how personal autonomy, even in a liberal society, cannot provide an unlimited scope for self-authorship. The distinction between hard and soft paternalism is not as firm as philosophers have assumed. The personal freedom granted by liberal respect extends only so far as is tolerated by the societal and institutional concepts of personhood and mental competence. A policy of hard paternalism, i.e. an attempt to place limitations *upon* personal autonomy, is better understood as a claim about the proper function of liberal personhood, limiting a person's freedoms from *within* the concept of personal autonomy.

2. Self-worth as a function of liberal personhood.

Liberal respect is an egalitarian principle: the claim that my attribution of normative authority is not superior to your own, with regard to your self-oriented choices. I may

show that you are wrong by referring to some mutually recognised authority, or by pointing to your own values, but I cannot simply assert the authority of my own intuitions over yours. Whether we understand this to be an expression of the neo-Millian concept of personal sovereignty (Feinberg 1984, 3:64–70; Arneson 1980; VanDeVeer 1986, 62; Kleinig 1983, 23), or of Kantian respect for another's capacity for giving and recognising reasons (Brock 1988; Scoccia 1990; Rawls 1971, 248–250), its aim is to allow people to engage in meaningful self-authorship, through the development and pursuit of their own conception of the good life. The challenge that this presents to 'hard' paternalism, i.e. interference with another's exercise of personal autonomy (Scoccia 2008, 351), is not that we suppose liberty to be a good of near-infinite worth, capable of outweighing any harm that the paternalist wishes to guard against. Rather, it is that such paternalism involves the imposition of values that are foreign to the person's self-conception (from 'Mental competence and its limitations'). In the context of suicide-prevention, this means demonstrating that the person *ought* to value her life (or some other value contingent upon life, such as her future happiness), even though this is inconsistent with her actual values.

This is no small imposition. In the context of the right to refuse life-saving treatment, respect for a person's decision to die is seen as central to, and perhaps even necessary for, the development and meaningful pursuit of that person's own conception of the good life. Whilst it is more of a declaration of shared principles than a comprehensive argument, 'The Philosopher's Brief' (Dworkin et al. 2007), by a remarkable collection of leading philosophers, provides possibly the clearest statement of the common ground shared by moral philosophers writing on this issue:

"*A person's interest in following his own convictions at the end of life is so*

> *central a part of the more general right to make 'intimate and personal choices' for himself that a failure to protect that particular interest would undermine the general right altogether. Death is, for each of us, among the most significant events of life." (Dworkin et al. 2007, 342)*

Taken at face value, their claim suggests that paternalistic interference in a person's death has importance *beyond* the immediate consequences of that decision; to 'undermine the general right altogether', it must detract from her *earlier* pursuit of autonomy, even if that pursuit was (up until now) successful. To make sense of this claim, we have to understand the meaningful pursuit of autonomy as turning upon the character of a person's life as a whole. Our most central goals and values are often not pursued through atomistic choices, but through longer-term projects that run as ongoing threads through our personal narratives (Meyers 2005, 37–38; O'Neill 2002, 83–94; Freidman 2005; Oshana 2005, 84–94). As such, the meaning of of our earlier efforts and sacrifices in the pursuit of such a project, i.e. whether they have been worthwhile or whether we see ourselves as having wasted that part of our lives, is often determined by the choices we make much later in that project. As the ultimate culmination of these projects, our choices in death are central to the character and shape of our overall life (Dworkin 1993, 230; Ott 2009; DeGrazia 2005, 71–84; Schechtman 1996, 94–114; Kuczewski 1994; Kuczewski 1999). As Dworkin puts it,

> *'There is no doubt that most people treat the manner of their deaths as of special, symbolic importance: they want their deaths, if possible, to express and in that way vividly to confirm the values they believe most important to their lives.' (Dworkin 1993, 211)*

In the accounts quoted above, the decision to die is a *continuation* of the projects that give meaning to one's life. Yet, as Velleman (1999, 611) observes, it makes no sense to values these projects unless one also values the life to which the projects matter. For Dworkin, there is no contradiction here: self-worth is not measured by our attitude towards our current or future survival, but by out attitude towards our life as a whole. The choice to die protects the value of the person's life as a whole by allowing her personal narrative to end in a manner that reflects her goals and values. In this sense, the decision to die can be strangely life-affirming, in that it affirms *a particular kind of life* to which the person attributes great value. For example, if independence and self-sufficiency are important to a person's conception of the good life, then her refusal of life-saving treatment in the face of rapidly advancing dementia may affirm the value that she attributes to having lived a life that is characterised by these qualities. Similarly, a Jehovah's Witness's religious refusal of blood transfusions is not a rejection of self-worth, but a continuation of the religious goals that she believes matter most to the value of her overall life.

This affirmation of self-worth goes some way towards capturing the ever-opaque concept of 'dying with dignity' that seems to pervade moral discussion about the choice to die. Like dignity, it implies both the authentic expression of one's character and also one's acknowledgement of some value that transcends one's subjective desire and experiences. Combining these two qualities, dignity demands an expanded self, that extends to include identification with principles that go beyond one's current survival and happiness. This is why the notion of dignity holds such gravitas in discussions of death; on an expanded account of self, the sacrifice of one's life can be an affirmation of the principles that one identifies with.

For Velleman this weighing of one's life against one's other interests is wrongheaded:

> *"But the dignity of a person isn't something that he can accept or decline, since it isn't a value for him; it's a value in him, which he can only violate or respect. Nor can it be weighed against what is good or bad for the person. As I have argued, value for a person stands to value in the person roughly as the value of means stands to that of the end: in each case, the former merits concern only on the basis of concern for the later. And conditional values cannot be weighed against the unconditional values on which they depend. The value of means to an end cannot overshadow or be overshadowed by the value of the end, because it is already only a shadow of that value, in the sense of being dependent upon it. Similarly, the value of what's good for a person is only a shadow of the value inhering in the person, and cannot overshadow or be overshadowed by it." (Velleman 1999, 613)*

Velleman's premise is that there cannot be value in a person's good, unless there is value in the person herself. From this, Velleman reasons that a person cannot reasonably reduce her lifespan *merely* in order to make it better as a whole, if that would violate the value that is inherent in *her*, rather than in her *good*. Velleman's error is to equate the person herself with that person's immediate survival as a rational being. Mere survival is neither unique to any one person, nor to personhood in general, and Velleman explicitly recognises this when he later concedes that we do not respect the value in a person by forcing her to live once unbearable pain has scattered her personality and impaired her rational capacities (Velleman 1999, 618). Nor can we so easily separate a person's self from that person's authentic values and convictions. As biological entities we may be defined entirely in terms of our survival and capacities,

but as *persons* – moral and social beings, whose subjective experience of agency is socially authenticated through the attribution of moral responsibility and autonomy – our self extends beyond survival to include the expanded features of our self-conception: our principles, cultural location, deeply-held values and life projects (Wolf 1989, 141; Meyers 2005; Velleman 2005, 65; Oshana 2005, 78; Quante 1999; Christman 2005, 345; Widdershoven and Berghmans 2001, 96). As DeGrazia notes, as *persons*, rather than mere organisms, we are inseparable from our temporal, social and psychological dimensions:

> *'We relate to each other not as momentary beings but as individuals with ongoing histories…it is especially important in the context of family ties, friendships, and other close relationships. Here we not only assume that others have personal narratives, which involves memories and plans, among other things (an assumption we make with strangers as well); we also take the narratives into account in our interactions with familiar individuals. And, in doing so, we attribute to each individual a personality and character, which involve relatively enduring traits. Such traits are psychological features of an individual.' (DeGrazia 2005, 21)*

Consider a scenario where the only way of surviving is by undergoing an organ transplantation, but the only viable source of an organ donation is to obtain the organ by grossly exploiting an impoverished youth. Quite separately to our concern for the youth, we have an excellent *self*-interested reason to refuse the transplantation, in that we do not want our lives to be so radically re-characterised by injustice. There is nothing self-repudiating in this choice. Our ongoing experiential interests are contingent upon our survival, but we also identify with moral and ideological

principles. In the context of the non-experiential values that go towards comprising the dignified self, Velleman mischaracterises the nature of the decision. We do not weigh the value in ourself against something that has value to ourself. By pursuing our deeply held non-experiential interests, we respect the value in ourself over the experiential interests that turn upon survival.

Nonetheless, Velleman's underlying claim has merit, at least as elucidating a view that is common within popular discourse on suicide: that self-worth, and the question of whether one *ought* to have self-worth, are not governed by our subjective desires. As I noted in the previous section, when faced with a person's denial of self-worth, there is a sense in which concern for her autonomy seems to miss the point – it wrongly suggests that self-apathy and suicidal depression are attitudes that are *capable* of warranting our respect, as though we would be obliged to defer to them if only the person was sufficiently mentally capable. This notion, that there is something morally and personally wrong with the denial of self-worth, that fundamentally excludes it from being a proper object of liberal respect, is contained within the psychiatric concept of depressive mental illness (American Psychiatric Association 2000, Appendix B).

Appealing to the self-affirming qualities of the choice to die allows us to accept Velleman's central claim – that there is a value *in* being a person, also known as dignity, that is independent of our subjective wants – while also providing that the choice can sometimes be reasonable. 'Dignity' in this sense is not reducible to autonomy (we cannot choose to exclude it through our self-authorship, no matter how mentally unimpaired and informed we are), but its satisfaction is intimately defined by the same goals and values that matter to personal autonomy. Conversely, a decision to

die can be voluntary, and free from mental impairment, yet violate dignity through its rejection of personal self-worth. As I noted earlier, liberal discussion of the right to refuse life-extending treatment typically addresses only certain kinds of decision to die: the right to refuse treatment is defended by reference to the elderly, terminally ill, grossly incapacitated or ideologically driven. Suicide amongst the young, healthy and opportunity-rich, in particular, is more commonly associated with the *denial* of self-worth, in expression of despair and hopelessness.

Velleman's positing of a categorical value in oneself shares much with the popular response to suicidal depression, i.e. that there is something wrong with this as an attitudinal state, not because it is inauthentic but because the person's values are wrong. There are at least some circumstances in which a person has value *regardless* of whether she subjectively recognises or identifies with that value. This is the central claim in Velleman's 'A right of self-determination?', from which I have been drawing at such length. Formally, Velleman's account adopts a Kantian position (2005, 611), but underlying this is a desire to develop a secular version of the religious appeal to a value in human life that extends beyond any person's capacity to choose:

> *'That's what I miss in so many discussions of euthanasia and assisted suicide: a sense of something in each of us that is larger than any of us, something that makes human life more than just an exchange of costs for benefits, more than just a job or a trip to the mall. I miss the sense of a value in us that makes a claim on us – a value that we must live up to.'* (Velleman 1999, 612)

It is striking how closely Velleman's appeal to a value greater than us, that we violate by shortening our life for the sake of making it better to us, resembles the concept to

which Dworkin appeals in defending the choice to die. Dworkin (1993, 201) posits the existence of 'critical interests', interests that we believe matter to us independently of whether we personally recognise them, such that our life would be genuinely worse off if we failed to acknowledge and pursue them. For Dworkin, however, our interest in pursuing these critical interests underlies the appeal to personal integrity. By identifying with goals and values that we believe matter to our flourishing, beyond our personal decision to value them, we are drawn to demand freedom to formulate and pursue our conception of these critical interests (Dworkin 1993, 216–224). For both Velleman and Dworkin, the purpose of liberal freedoms is, in part, to live up to these greater values, whether by respecting the intrinsic worth of our own life, or by seeking to pursue one's deeply held goals and values in death.

This same sense of dignity, or greater meaning in personhood, is missing in the traditional concept of mental competence, with its focus upon the procedural capacities required for decision-making. Here is a typical definition of this traditional conception of competence, still dominant in Bioethics, as expressed in a highly influential text by Beauchamp and Childress:

> *'normal choosers who act (1) intentionally, (2) with understanding, and (3) without controlling influences that determine their action...it needs only a substantial degree of understanding and freedom from constraint'*
> (Beauchamp and Childress 2009, 101)

As I explain in 'Mental competence and its limitations', mental competence is a necessarily evaluative concept. Moral responsibility and autonomy are not qualities that we are capable of obtaining (at least not to the kind of extent that our moral

expectations presume), but statuses that we attribute in accordance with our evaluative beliefs regarding the proper function of personhood in our society. Viewed from this perspective, the traditional conception of mental competence, with its emphasis upon the mental capacities required for the basic processes of decision-making, attributes to personhood a minimal, nearly non-existent function. We are competent (and hence functional) insofar as we are capable of effectively pursuing our wants, with no restriction beyond the immediate authenticity of our current want and our capacity to process and apply the information relevant to it (see 'Mental competence and its limitations'). As such, the value inherent in the corresponding concept of personal autonomy is nothing more than the consequential benefits of free choice, always questionable as a basis for a broad policy of restricting paternalism (Wrigley 2007, 387–388; Valerius 2006, 455–456; Meyers 2005, 47–48), let alone in a medical context in which decisions concern complex scientific information and expert judgment, where the patient is removed from her usual environment and subjected to unfamiliar institutional processes.

As I have argued in 'Reasons, autonomy and paternalism' and 'Mental competence and its limitations', the liberal construction of personhood is properly concerned with self-authorship, rather than freedom over a series of atomistic choices. This account constructs the function of personhood through the claim that the 'good life' for a person is one lived in accordance with one's personal integrity. As discussed at length in 'Reasons, autonomy and paternalism', the appeal to personal integrity has a basis in liberal accounts of autonomy (Feinberg 1984, 64–70; Arneson 1980; VanDeVeer 1986, 62; Kleinig 1983, 23) and practical reason (Brock 1988; Scoccia 1990; Rawls 1971, 248–250; O'Neill 2002, 91), whereby the self is the ultimate authority over our self-oriented interests and our reasons for action respectively.

However, a goal of self-authorship, by itself, does not imply dignity, in the sense of finding meaning in oneself that transcends one's personal attribution of value. If the good life is to be determined by self-authorship, why we should we exclude particular attitudes, i.e. the denial of self-worth, from those that we may authentically hold? Why would it be any less an act of self-authorship to identify with goals and values that are self-loathing and self-destructive in nature? More broadly, why should we suppose that 'dignity' holds any special place at all within our self-authorship on matters concerning the value of personhood? Why couldn't we authentically adopt goals and values that are not wanton desires, but are nonetheless self-centred in that we attribute worth to our goals and values, *simply because we value them*, with no claim to any personal worth that transcends self-interest?

An inseparable feature of the manner in which we authentically held goals and values is that we attribute to them a normative force. This claim that our goals and values hold normative authority, at least with regard to our own flourishing as a person, is fundamental to both the Millian and Kantian constructions of the claim that we can have reason to act against our own well-being 'see 'Reasons, autonomy and paternalism' and 'Mental competence and its limitations'). If we strip away the appeal to universals of human flourishing, we lose this normative component. For me to truly commit myself to a goal, in the sense of holding it to have real value *as a value*, I must see a good in it that is not reducible to my subjective desire; if I am truly committed to the goal of, say, finding meaning in music, I must consider that goal to have a value that cannot be fully and qualitatively replaced by a commitment to painting, even if I value both equally (or else these are just instrumental means of pursuing some further goal, such as artistic satisfaction, or happiness). The goal holds *normative* value to us,

because we value it for its own sake, such that we believe we would be genuinely worse off by failing to pursue it, even if we were equally happy. Dworkin recognises this in his discussion of critical interests:

> *'When we reflect about what makes a person's life good, we are torn between what seem to be antagonistic beliefs. On the one hand a person's critical interests seem very much to depend on his personality...But that appealing idea is hard to reconcile with an even more fundamental conviction we also have – that a person's thinking a given choice right for him does not make it so, that the sometimes agonized process of decision is a process of judgment, not just choice, that it may go wrong, that one may be mistaken about what is really important in life. That belief is indispensible to the basic distinction between critical and experiential interests, and to the challenge and tragedy most people feel.'* (Dworkin 1993, 206)

This, in itself, goes part-way to illustrating the sense of tragic waste that we associate with 'undignified suicide'. There is something terribly shallow about committing oneself to self-annihilating goals; not in the sense of 'shallowness' as a person who cares about very little, but in the sense that the person eschews the possibility of a life in which she could find value and meaning that is not centred entirely around her own wants. Nonetheless, the quality of dignity in liberal personhood lies not in the commitment to attributing value to our goals and projects, but in the process by which we develop the personal identity that makes self-authorship possible. The reality of self-authorship is a haphazard, even accidental, process. Nobody can design their own character from the ground up, nor is the process of developing one's goals and values strictly distinct from the practices through which we pursue those goals and values. If

one had to choose a solitary and exclusive design metaphor, self-authorship is closer to sculpture than narrative. We can only build upon the self that our cultural location, history and biology provides us with. Oshana, for example, describes how she could not fully exclude those aspects of her character that have arisen from her experiences as a black woman in America without considerable inauthenticity (2005, 83–91), and I would add to that the converse, that no amount of reflective deliberation would permit me to authentically adopt those aspects of Oshana's self. Similarly, to describe self-authorship as consisting entirely, or even mostly, of reflective choice understates the importance of self-discovery.

Personal identity (in the sense of character) is not something that we neatly design and choose, but is instead a process of experimentation, playful trial and discovery within our cultural and biological boundaries. Dworkin, in particular, notes the experimental qualities of self-authorship where, faced with the impossibility of rationally constructing an identity as a wholly pre-designed exercise, he recognises the importance of discovery *through* self-authorship:

> *'People do not make momentous decisions like these by trying to predict how much pleasure each choice might bring them. We sometimes say that we discover our own identities through such decisions. But we do not mean that we discover how we have already been wired...We can search our past for clues about what satisfies us or makes us happy, but life-shaping decisions are also occasions for imaginative fantasy and, above all, for commitment.' (Dworkin 1993, 205)*

The more important choices in our life – e.g. beginning or ending relationships,

marriage, to start or change a career, to migrate between countries, or whether to have offspring and how to raise them - do not permit exhaustive foreknowledge. Time and time again, we plunge into our commitments and only later decide whether we identify with the resulting way of life. For example, upon reaching young adulthood we choose a course of study or employment, expecting it to govern our career, financial means and those aspects of life which turn upon finance and status, informed by nothing more than a hazy combination of advice, role models and generalised expectations. We often begin both friendships and intimate relationships on the basis of chance circumstances, discovering the important details of the relationships as they go. We may or may not choose whether to have children, with either choice reflecting at best an educated guess as to the effect upon our future way of life. In each case, we identify with the *desire* to make the choice, and sometimes with our expectation of how the choice will turn out, but it is only afterwards that we get the opportunity to know and identify with the full relevance of the choice upon our way of life. All the while, there remains the sneaking suspicion that it is our choices that influence our priorities and character, rather than our self that has determined the way of life we have developed.

Without this quality of experimentation, and the personal growth that it permits, we could not develop our own identity. At no stage do we ever have the opportunity to construct an identity in isolation from the pursuit of our lives. Rather, we experiment with changes to our self and to our way of life, plunging in and out of new roles, and developing our self as we go; hoping that if our lives go well we may steadily modify our self into something capable of giving meaning to our lives. Through this experimentation we find values through which we imbue our lives with normative worth. Most importantly, the values that motivate our life choices are not formed *prior* to our pursuit of personal autonomy. The process by which we find meaning and worth

in our lives is the same process by which we develop a sense of who we are. Without the search for meaning and worth in our lives, we could not have an authentic identity to express through the exercise of self-authorship. Liberal personhood is a constant and active search for meaning.

With this capacity for personal growth in mind, we can begin to describe what is so badly dysfunctional about a person who is young and opportunity-rich but denies all self-worth. We do not call the person dysfunctional because we dispute the quality of her subjective unhappiness. In fact, the more authentic and severe her denial of value, the more dysfunctional we take her to be. We may suspect that the person is failing to recognise her own intrinsic worth, insisting that her denial of self-worth is factually wrong, and implying that she *ought* to value herself even if she doesn't. However, if we have liberal sympathies, we will not want to endorse the proposition that some individuals should be denied equal status in attributing normative authority, simply because we disagree with their values. This person might not hold any special advantage in judging whether she intrinsically has worth (Velleman 1999, 611), but then again, neither do we. Besides, many of us are convinced that there are circumstances where it is perfectly reasonable to say that one's life lacks value, and that these circumstances seem to be intimately connected to the person's own attribution of worth (such as where the person is no longer capable of meaningful identity or personhood in due to the combination of mental, physical and social restrictions that they face). No, she is dysfunctional because she is no longer *searching* for value and meaning – of any kind. Repudiating the values that might have given meaning to her life is a decision requiring our respect, but this individual has opted out of the process of personhood itself. Personhood is not just the holding of a particular set of goals and values against which one judges the worth of one's life. We have a

temporal component, through which we pursue projects and construct personal narratives over time (Schechtman 1996, 94–126; DeGrazia 2005, 78–107; Quante 1999, 365–381; Quante 2007, 56–76; Kleinig 1983, 67–68; Edwards 2010). A critical component of personhood is the process by which people search for value through different permutations of their identity over time, responding to adversity through personal reinvention. To have periods in which one's current goals and values fail to bring meaning to one's life can be part of functional personhood, but to cease searching for value and meaning is dysfunctional. By abandoning the process of personal growth, she has repudiated her own personhood.

3. Limitations upon the liberal self

Once we are no longer able to hide behind the claim to normative objectivity in the categorisation of mental illness and mental competence, we cannot simply defer to the psychiatric classifications of mental illness in determining whether a particular condition should deny personhood. Instead, we need a robust positive account of liberal personhood, capable of meaning and value that can be justified on its own terms, rather than as a pseudo-factual claim about the mental capacities needed for competent decision-making. The basis upon which we attribute mental responsibility and autonomy, the cornerstones of personhood in our moral reasoning, must itself be grounded in an account of the good.

Under the account that I have developed in 'Reasons, personhood and autonomy' and 'Mental competence and its limitations', that good is constructed in terms of an appeal to personal integrity, under the belief that the person's integrity – i.e. her development and pursuit of conception of the good, in a manner authentic to her own personal identity – determines the good life for that individual. The purpose of liberalism, on

this view, is not to protect a variety of equally legitimate conceptions of the good, under the pretence that they are all equally true, or in denial of the possibility of universal facts about human or individual flourishing. It is to protect the attempts by individuals, who deserve morally and socially equal status in society due to their shared status of personhood, to pursue their own conceptions of what human flourishing entails. We might, under a neo-Millian account of the manner popularised in the 1980s by Feinberg (1984, 65–70) and VanDeVeer (1986, 62), take this to rest upon the claim that self-authorship is itself a part of the universal requirements for human flourishing. Alternatively, we might ground this principle in respect for persons, acknowledging that in a matter so important as a person's critical interests, we lack moral standing to impose upon him our conception of what critical interests he should follow, unless we can point to some instrumental error or mental inefficiency on his part (Scoccia 1990; Rawls 1971, 248–250; O'Neill 2002, 91).

In giving a positive account of liberal personhood, the more that we seek to imbue that concept of personhood with meaning beyond the individual, the more restrictive that account of personhood becomes. If we understand the function of personhood to be nothing more than the processing and applying of information in regard to atomistic questions, we would have no basis to exude particular attitudes or motivations (such as the denial of self-worth) from the scope of mental competence. Conversely, if we take the purpose of personhood to be self-authorship, in the construction of a robust self comprised by the ideals, goals and values that we identify with, we require capacities that go beyond our own mental abilities: we need social relations (Meyers 2005, 33–35), an environment that does not grossly restrict our options for personal development on the basis of personal impairments (Oshana 2005, 85–94) and, as I argue in this paper, the affirmation of self-worth.

Moreover, on a proper understanding of self-authorship, it is not so much a process of design but as experimentation, where we try on different social roles, weather the inevitable bombardment of events beyond our control, and in the process seek to formulate an identity through which we may find meaning in our lives. This takes us a very long way from the absolute freedom of informed consent that dominates bioethics, and I can understand why many would find this more vulnerable to encroaching authoritarianism. After all, how much difference can there be between overriding a competent and informed person's authentic values on the basis of hard paternalism, and imposing the same involuntary treatment by holding that the person's resistance to treatment as evidence of mental incompetence? However, while the distinction between the capacities and ends of personal autonomy have been softened, they have not been dissolved altogether. By reconstructing a moral claim that people ought to put aside their authentic values in the pursuit of a wholly external motivation, into a claim that the value is a criterion for mental competence, we transform our concern from a concern about morality into a concern for human flourishing. This provides a framework through which we may engage such claims, avoiding the often fruitless dispute over our differing intuitions about the importance of freedom and well-being. We can ask how these restrictions relate to the proper function of personhood, and (under a liberal conception of personhood) evaluate them in the context of the restrictions required for meaningful self-authorship.

At the most immediate and practical level, we must realise that as soon as we accept that mental competence is an evaluative concept, we can no longer maintain a blanket principle against hard paternalism. Mental illnesses alienate mental qualities from our personal identity and in doing so they place absolute restrictions upon the permissible

scope of self-authorship. The more robust account of self-authorship given in 'Mental competence and its limitations' necessarily implies a greater degree of hard limitations upon the scope of paternalism, as we distinguish between a *meaningful* exercise of autonomy (one that respects dignity or self-worth, or one that requires social relations conducive to the relational competent of autonomy) and mere freedom of unstructured choice.

By softening the distinction between hard and soft paternalism, and openly acknowledging that the evaluative nature of mental competence may limit legitimate self-authorship beyond the procedural requirements of voluntariness, we invite the criticism that we endorse the encroaching subjugation of the individual to social norms under the guise of the requirements for 'meaningful' autonomy. I have already made the case that the evaluative nature of mental competence is routinely abused in this way, in 'Mental competence and its limitations'. If 'genuine' autonomy requires the patient to comply with socially or institutionally mandated standards of competence, it may invite the charge that we place the patient completely at the mercy of the society from which we are demanding her freedom.

My first response to this is that we have no other choice. Competence *is* evaluative, and hiding that only makes the resulting injustice worse for its invisibility (see 'Mental competence and its limitations'). But secondly, Tailor's (1985, 215–216) defence of positive liberty comes to mind. The risk of being defeated by authoritarian conceptions of personhood is no reason for us to adopt an inaccurate, but convenient, pretence of value-neutrality. We must meet the authoritarian argument directly, rather than seeking to avoid it by positing a neutral concept of autonomy that has never truly existed. Rather than pretending that the informed consent doctrine can provide meaningful

freedom by refraining from normative intervention, we must defend a positive account of liberal personhood from which a basis for patient autonomy can be developed.

References

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Arlington, Virginia: American Psychiatric Publishing. http://dsm.psychiatryonline.org.ezproxy.library.uwa.edu.au/book.aspx?bookid=22.

Arneson, Richard. 1980. "Mill versus Paternalism." *Ethics* 90 (4): 470-489.

Beauchamp, Tom, and James Childress. 2009. *Principles of Biomedical Ethics*. 6th ed. New York: Oxford University Press.

Brock, Dan. 1988. "Paternalism and Autonomy." *Ethics* 98 (3): 550-565.

———. 1992. "Voluntary Active Euthanasia." *The Hastings Centre Report* 22 (2): 10-22.

Buchanan, Allen, and Dan Brock. 1989. *Deciding for others: the ethics of surrogate decision making*. Cambridge: Cambridge University Press.

Cantor, Norman. 1993. *Advance Directives and the pursuit of death with dignity*. Medical Ethics. Indianapolis: Indiana University Press.

Christman, John. 2005. Autonomy, Self-Knowledge and Liberal Legitimacy. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson. Cambridge: Cambridge University Press.

DeGrazia, David. 2005. *Human Identity and Bioethics*. New York: Cambridge University Press.

Dworkin, Ronald. 1993. *Life's Dominion*. New York: Knopf.

Dworkin, Ronald, Thomas Nagel, Robert Nozick, John Rawls, Thomas Scanlon, and Judith Jarvis Thomson. 2007. The Philosopher's Brief. In *Bioethics: Introduction to History, Method and Practice*, ed. Nancy Jecker, Albert Jonsen, and Robert Pearlman. 2nd ed. Sudbury, Massachusetts: Jones & Bartlett Publishers.

Edwards, Craig. 2010. “Beyond Mental Competence.” *Journal of Applied Philosophy* 27 (3): 273-289.

Feinberg, Joel. 1984. *Harm to Self*. Vol. 3. 3 vols. The Moral Limits of the Criminal Law. New York: Oxford University Press.

Freidman. 2005. Autonomy and Male Dominance. In *Autonomy and the Challenges to Liberalism*, ed. Christman and Anderson, 150-173. Cambridge: Cambridge University Press.

Harvey, Martin. 2006. “Advance Directives and the Severely Demented.” *Journal of Medicine and Philosophy* 31: 47-64.

Jansen, Lynn. 2006. “Hastening Death and the Boundaries of the Self.” *Bioethics* 20 (2): 105-111.

Kadish, Sanford. 1992. “Letting Patients Die: Legal and Moral Reflections.” *California Law Review* 80: 857-888.

Kihlbom, Ulrik. 2008. “Autonomy and negatively informed consent.” *Journal of Medical Ethics* 34: 146-149.

Kirby, Michael. 1983. “Informed consent: what does it mean?” *Journal of Medical Ethics* 9: 69-75.

Kleinig, John. 1983. *Paternalism*. Totowa: Rowan and Allenheld.

Kuczewski, Mark. 1994. “Whose will is it, anyway? A discussion of advance directives, personal identity and consensus in medical ethics.” *Bioethics* 8 (1): 27-48.

———. 1999. “Commentary: Narrative Views of Personal Identity and Substituted Judgement in Surrogate Decision Making.” *Journal of Law, Medicine and Ethics* 27 (1): 32-36.

McMahan, Jeff. 2002. *The ethics of killing: problems at the margins of life*. New York: Oxford University Press.

Meyers, Diana Tietjens. 2005. Decentralizing Autonomy: Five Faces of Selfhood. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 27-55. Cambridge: Cambridge University Press.

O'Neill. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

Oshana. 2005. Autonomy and Self-Identity. In *Autonomy and the Challenges to Liberalism*, ed. Christman and Anderson, 77-97. Cambridge: Cambridge University Press.

Ott, Andrea. 2009. "Personal Identity and the Moral Authority of Advance Directives." *The Pluralist* 4 (3): 38-54.

Quante. 1999. "Precedent Autonomy and Personal Identity." *Kennedy Institute of Ethics Journal* 9 (4): 365-381.

———. 2007. "The Social Nature of Responsibility." *Journal of Consciousness Studies* 14 (5-6): 56-76.

Rawls. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Savulescu. 1997. "The trouble with do-gooders: the example of suicide'." *Journal of Medical Ethics* 23: 108-115.

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornwell University Press.

Scoccia, Danny. 1990. "Paternalism and Respect for Autonomy." *Ethics* 100 (2): 318-334.

———. 2008. "In Defense of Hard Paternalism." *Law and Philosophy* 27: 351-381.

Shatzberg, Alan, Jonathon Cole, and Charles DeBattista. 2010. Emergency Department Treatment. In *Manual of Clinical Psychopharmacology*. 7th ed. Arlington, Virginia: American Psychiatric Publishing. http://psychiatryonline.org.ezproxy.library.uwa.edu.au/content.aspx?bookid=2

&sectionid=1363762#608288.

Simon, Robert. 2008. Suicide. In *The American Psychiatric Publishing Textbook of Psychiatry*, ed. Robert Hales, Stuart Yudofsky, and Glenn Gabbard. 5th ed. Arlington, Virginia: American Psychiatric Publishing. http://psychiatryonline.org.ezproxy.library.uwa.edu.au/book.aspx?bookid=3.

Taylor, Charles. 1985. What's wrong with negative liberty? In *Philosophy and the Human Sciences*, 211-229. Philosophical Papers 2. UK: Cambridge University Press.

Valerius, Jukka. 2006. "On Taylor on Autonomy and Informed Consent." *The Journal of Value Inquiry* 40: 451-459.

VanDeVeer, Donald. 1986. *Paternalistic Intervention: The Moral Bounds on Benevolence*. New Jersey: Princeton University Press.

Velleman. 2005. The Self as Narrator. In *Autonomy and the Challenges to Liberalism*, ed. Christman and Anderson, 56-76. Cambridge: Cambridge University Press.

Velleman, James David. 1999. "A right of self-termination?" *Ethics* 109 (3): 606-628.

Weigel, Margaret, David Purselle, Barbara D'Orio, and Steven Garlow. 2009. Treatment of Psychiatric Emergencies. In *The American Psychiatric Publishing Textbook of Psychopharmacology*, ed. Alan Schatzberg and Charles Nemeroff. 4th ed. Arlington, Virginia: American Psychiatric Publishing. http://psychiatryonline.org.ezproxy.library.uwa.edu.au/content.aspx?bookid=29&sectionid=1365913#425748.

Widdershoven, and Berghmans. 2001. "Advance directives in psychiatric care: a narrative approach." *Journal of Medical Ethics* 27: 92-97.

Wolf. 1989. Sanity and the Metaphysics of Responsibility. In *The Inner Citadel: Essays on Individual Autonomy*, ed. Christman, 137-151. New York: Oxford University Press.

Wrigley, Anthony. 2007. “Personal identity, autonomy and advance statements.” *Journal of Applied Philosophy* 24 (4): 381-396.

Paper 5 - Beyond Mental Competence

Abstract

Justification for psychiatric paternalism is most easily established where mental illness renders the person mentally incompetent, depriving him of the capacity for rational agency and for autonomy, hence undermining the basis for liberal rights against paternalism. But some philosophers, and no doubt some doctors, have been deeply concerned by the inadequacy of the concept of mental incompetence to encapsulate some apparently appealing cases for psychiatric paternalism. In this paper, I continue my argument that we ought to view mental incompetence as just one subset of a broader justification for psychiatric paternalism. I do this by identifying a group of mental states that are widely associated with dubious personal autonomy, yet cannot be accounted for by the traditional concept of mental competence. The features of these mental states that render them dubious with regard to personal autonomy concern the manner in which the individual *changes*; i.e. they are to do with the difference between an individual's authentic character and the qualities imposed by mental illness.

In 'Mental competence and its limitations' I conceded that, to be maximally effective, my account required some theoretical basis for distinguishing the qualities imposed by illness from a person's 'authentic' personal identity, upon which the appeal to personal integrity can be based. In the course of addressing the phenomenon of 'judgment shift' – change imposed by mental illness that prevents meaningful autonomy despite the exercise of informed and mentally capable judgment – I establish such an account of personal autonomy, at least in terms of an initial framework to guide further development. As such, in addition to addressing the applied problem at the centre of the paper's concerns, it plays a dual role in furthering my broader thesis. Through my account of personal identity, I further my claim that involuntary treatment is justified

essentially by social and moral norms, while accepting a concept of mental illness where such illness is both real and deserving of treatment.

Beyond Mental Competence

1. Introduction: Self and psychiatric paternalism

Psychiatric paternalism is paternalism aimed at preventing a person self-harming (not just physically) due to mental illness. Justification for psychiatric paternalism is most easily established where it renders the person mentally incompetent, depriving him of the capacity for rational agency and for autonomy, hence undermining the basis for liberal rights against paternalism. But some philosophers, and no doubt some doctors, have been deeply concerned by the inadequacy of the concept of mental incompetence to encapsulate some apparently appealing cases for psychiatric paternalism (Culver and Gert 1990; Davis 2008; Quante 1999) Should we intervene to save a patient who is suicidal due to a readily treatable depressive condition, but is not psychotic, has neither delusional symptoms nor substantial cognitive impairment, and is mentally capable of understanding and reasoning about the factors relevant to that decision? Or what about the infamous case of Joyce Brown, who was found by the US Supreme Court to be mentally competent, yet to have also suffered drastic deterioration in living conditions, apparently due to the change in character brought on by schizophrenia (Failer 2002, 2–28)?

Such cases are not without controversy. In the Joyce Brown case, successive courts diverged on whether to refuse involuntary treatment because of her mental competence, or to order such treatment because of her character change and subsequent isolation and homelessness. The case was never fully resolved (see endnote for a brief history[v]). Nonetheless, I suspect that at least a substantial minority of philosophers, and a majority of psychiatrists, would approve of intervention. I believe that paternalistic intervention is warranted, but to describe such cases as examples of

mental incompetence would be to stretch the term such that it loses its usefulness as a non-circular justification for paternalism. Instead, we ought to view mental incompetence as just one subset of a broader justification for psychiatric paternalism. The very basis of liberal limitations on psychiatric paternalism, whether described in terms of rights to autonomy or as respect for differences in values and lifestyles, presupposes a sense of moral persistence and hence some sufficiently persistent self. Paternalistic intervention is warranted when we are unable to govern our lives in a manner consistent with the goals and values that comprise that 'self'. One way that can occur is when we lack the mental capacities required for autonomy, such that we are unable to interpret and interact with our environment in order to meaningfully pursue our goals, i.e. mental incompetence. But it can also occur when we are subject to impositions that alter our goals without altering our capacity to pursue them – i.e. when it is our 'self' that is impaired rather than our competence.

This impairment of self, or *'judgment shift'* as I will refer to it, is perhaps most apparent in some examples of non-delusional mood disorders, but it is not limited to the strictly psychiatric sphere.[3] We are familiar with the notion of drug addicts acting 'out of character', not just when intoxicated, but at other times when driven by craving or even just affected in mood. Similarly, a person who is subject to great stress, or subject to indoctrination, may act against his usual goals and values, without being mentally ill or incompetent. In this paper I am concerned with an exceptional, but deeply troubling, subcategory of these phenomena. Often, when our goals and values

[3]. I recognise that addiction in particular is often viewed as a psychiatric illness (American Psychiatric Association, 2004, *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*, Washington DC: American Psychiatric Association). My point is that addiction and indoctrination are not phenomena that we *limit* to the psychiatric sphere. We address them simultaneously as social and moral concerns, as well as psychiatric phenomena.

change due to mental illness, indoctrination or addiction, we experience an inner conflict such that on reflection and with clear insight we see an emotional and cognitive conflict between the newly imposed goals and our more central or 'settled' goals[4]. There is already an influential strand of liberalism that denies that these (imposed) goals are truly autonomous, allowing that we should intervene to protect a person's more central goals and values from transitory or superficial desires (Kleinig 1983, 66–69). However, there are other, perhaps less common, examples where a person wholeheartedly adopts the newly imposed goals and values. In such cases there doesn't seem to be any 'inner conflict', or at least no more than in any other wholehearted choice. I am referring to the apparently willing addict, the passionately converted indoctrinated, and the mood disordered person who views her new way of thinking as an extraordinary revelation rather than a temptation to struggle against. The goals being imposed by these conditions are not *threatening* the bearers' more central goals and values – they *are* the bearers' most central goals and values. Yet prior to the imposition of addiction, illness or indoctrination, the person would have rejected those goals, and so we are tempted to view them as some disordered imposition rather than the constitution of the person's legitimate character.

The mere presence of mental illness does not by itself justify paternalism; we must instead show that the illness warps the person's decision-making in some relevant way. Neither does sudden or radical change: would we really wish to rule out sudden self-discovery, and if so, what basis could we find for doing so? The decisions in question appear to exemplify the kind of liberal autonomy upon which we base restrictions on paternalism. The people involved are either mentally competent, or at least not readily

[4]. I prefer the term 'central', as 'settled' implies that the imposed goals would gain legitimacy if only we were to refrain from intervention long enough.

proven to be incompetent. They are following the goals that they identify with and that are most central to their character. These are decisions that they would *endorse* (or at least fail to repudiate) if they were to reflect upon them. Yet the suddenness with which these goals have appeared, and the radical discrepancy between their past and present personality is disturbing in a way that many people feel *should* invite paternalism even from a liberal perspective.

In this paper, I argue that the focus on mental incompetence as a pre-requisite for psychiatric paternalism has hampered our theoretical and practical response to such cases. The value of liberal restrictions on paternalism lies in the protection of personal integrity, in the goals and values that comprise our character. Some liberals have long viewed this as providing a broader basis for paternalism, permitting our intervention when even a mentally competent person endangers the goals and values that are more central to that person's autonomy and integrity (Kleinig 1983, 67–69; Faden and Beauchamp 1986, 262–269). I suggest that we go a step further. We ought to view character as containing a temporal component, such that the same numerical individual may suffer a severance of character that alienates her from her prior choices. This may give us good reason to paternalistically *prevent* the person from acting upon what are now (following the severance of character) her central goals, instead protecting the goals that had comprised her character prior to that severance. I will refer to this claim as 'the other self thesis', reflecting the claim's central thesis that one's character has a temporal component, such that extraordinary events may sever a person from holding responsibility or authenticity regarding decisions made previously.

There is a common and very sensible rebuttal to a claim that a person has lost responsibility, autonomy, authenticity, authority or some similar attribute due to some

alteration in her personality: simply *that character change and growth is an ordinary aspect of human life*. There are, admittedly, some circumstances where it is natural to associate change with an alteration in responsibility – loss of responsibility for juvenile crimes, for example – but these are associated more with a growth in mental faculties than a change in personality. The kinds of character change that I have in mind are not analogous to the growth from youth to adulthood. But any theory that seeks to portray autonomy and associated concepts as being at odds with ordinary character change faces a remarkable burden. I hope to avoid that burden by showing that the kinds of change I am concerned with are not ordinary. There are morally meaningful differences in the changes caused by mental illness and indoctrination and those that occur in the ordinary course of character development. I hope to give these differences a philosophical underpinning, showing how a person may sensibly identify with a vastly different former self in ordinary circumstances, while being unable to do so where the changes between her former and latter selves are imposed by mental illness, addictions or similar conditions.

2. What is Mental Competence?

'Competence' is simply the state of being capable of performing a particular task effectively. The term's precision varies with the precision of the task that it refers to – for example, it is a far less precise description to call someone a competent swimmer than it is to say that someone is competent to swim the English Channel in clear weather. 'Mental competence' refers to the set of mental capacities that are required for a particular task. The term has gained a legal and medical significance associating it with the contexts of decision-making and paternalism, but outside of that context one could also be mentally competent for other tasks such as passing an exam or solving complex mathematical problems. Our current interest in mental competence is as a

guide to when paternalism is morally permissible. One might say, then, that we are concerned with the mental capacities required in order to have a moral claim to freedom from paternalistic intervention. No doubt the term 'mental competence' is used on occasion to mean precisely that. But such a notion of mental competence adds nothing to the analysis – it is just a round-about way of redescribing the question. It is more productive to talk about competence for the task of possessing some independent trait from which rights against paternalism can be derived. For this reason, in the context of psychiatric paternalism, 'mental competence' typically refers to one's capacity for meaningful autonomy, the foremost component of which is the capacity for rational agency.

The precise nature of autonomy is itself a major topic of philosophical inquiry, and it would not be fruitful to try to establish an authoritative definition here. Feinberg (1984, 3:31–44) provides an excellent, albeit imprecise, guide to the concept of autonomy, describing it as comprising a vast list of interrelated mental qualities including self-possession (being one's own person), distinct self-identity, authenticity (not being the mouthpiece of other persons), self-determination, self-legislation, moral independence, integrity, self-fidelity, self-control, self-reliance, self-responsibility and self-generation. Other accounts of autonomy have sought to place greater emphasis on our existence as social and relational beings (Meyers 2005) and on the cultural and social conditions that are required for autonomy (Oshana 2005, 77–97; Anderson and Honneth 2005, 127–149; Friedman 2005). I will assume that most readers will have some concept of what autonomy is, although there is likely to be considerable variation in the details of that concept.

The relevance of autonomy to paternalism arises from two different (but not mutually

exclusive) liberal theories. Millian liberalism champions an overriding human interest in pursuing a highly autonomous life, not necessarily in terms of solitude or demanding causal independence, but rather one in which people have the ultimate moral authority over the design and path of their personal lives (Mill 1977, 265–267; Feinberg 1984, 3:65–70; Arneson 1980; Kleinig 1983, 22–25; Archard 2008). Under those accounts, our interest in autonomy takes moral precedence over the avoidance of self-harm. This is not the all-encompassing declaration of liberty over safety that it is occasionally misinterpreted to be. The overriding interest is in autonomy, rather than choice per se, and so this interest is negated if one is incapable of exercising meaningful autonomy.

The other category of liberal theories of rights against paternalism are those that (a) permit paternalism where a person's choice involves unreasonable risk of self-harm (taking into account any value given to autonomy), but (b) determine the reasonableness of a person's actions by their relationship to the person's goals and commitments, conjoined with some theory that ranks the comparative authority of different goals by reference to their deep-held subjective importance (Boddington 1998; Scoccia 1990; Brock 1988; Rawls 1971, 248–250). The proponents of this view don't form an obvious philosophical substrain in the nature of their Millian counterparts, but the view's association with Rawls (1971, 248–250) and its principle of tying liberty rights to reasonableness might justify our describing it as a Kantian form of liberalism (Brock 1988; Scoccia 1990; Rawls 1971, 248–250). This tradition does not directly claim that the 'good life' is one of greatest autonomy, but nonetheless requires authenticity as a measure of prudential reasonableness. Rationality may be universal, but it is not disconnected from what is authentic for the individual, and what is *prudentially* reasonable, on this view, is determined by a person's more central goals

and values. However, it also requires that a person possess the capacities required for authenticity and autonomy if she is to reasonably choose to risk her well-being.

3. Competence, judgment shift and defending the liberal approach to paternalism

Competence has long been the primary focus of psychiatric paternalism. Mental competence is a pre-requisite for both of the major liberal bastions of rights against paternalism: the claim that we have a right to personal autonomy, and the claim that it is our own deep goals that determine what is reasonable for us to do. Mental *in*competence disrupts our self-determination and undermines our ability to form and pursue our goals. Having written that, a curious feature of systems of psychiatric paternalism is that they seek to protect only a specific class of people, i.e. that subset of the population who are mentally ill as defined by the enabling legislation. Such systems do not protect against incompetence 'in general', but only that which is caused by mental illness.

Unsurprisingly a great deal of debate has sprung up over whether other impediments to self-determination and, goal-satisfaction should also provide grounds for paternalism. If we can intervene to prevent a person's mental incompetence from undermining his goals, then why not intervene where a mentally capable person's stubborn misjudgement will have the same result (Kleinig 1983, 67)? Why not intervene to prevent others from reaching unwanted outcomes through gross errors of fact, or even to prevent them devoting time and resources towards following religions that we believe are absurd (or, from the theist's perspective, why shouldn't they enforce their religion to prevent others from going to hell)?

These are all legitimate issues of philosophical debate, but the problem which I wish to

discuss is a different matter altogether. The examples I raised in the introduction - of the mood disordered, the person whose character is distorted by great stress, the (non-intoxicated) addict and the indoctrinated or brainwashed – are not people who are *failing* to effectively form and pursue their goals. Instead, they are quite certain of their priorities and are mentally capable of pursuing them. Such people are not mentally incompetent, unless we broaden that term to include anyone who we think should be the subject of paternalism. Their choices reflect the shape that they wish their lives to take, they are self-authoring and their choices are reasonable relative to their deeply held goals. And yet they are deeply enticing as cases for paternalism.

I believe that the danger in failing to account for the appeal of these cases is double-edged. If (as I believe is the case) these *are* cases where paternalism is justified, we do great wrong to the people involved by failing to protect their 'legitimate' or 'real' characters and goals from the ravages of serious illness. However, despite arguing for an expansion of psychiatric paternalism beyond the mentally incompetent, I consider my views to fall squarely within the 'liberal' approach to paternalism. I fear that such cases greatly endanger that liberal approach unless a better explanation of them is given. The 'liberal approach' that I refer to is not to do with prohibiting psychiatric paternalism, or restricting it to the narrowest justification possible, but has to do with the *kinds* of justification that can be given for paternalistic intervention. I refer to an account of paternalism as liberal if it holds that paternalism can *only* be justified by reference to the person's *own* central goals, desires, values or other similar conative states. I call these justifications 'liberal' because they are plausibly described as promoting the person's autonomy and liberty. If asked, 'why should we stop such a person from pursuing her immediate intentions in order to protect some other goal?', the purveyor of the liberal approach can answer 'because he himself, in some morally

important manner, wants that goal to be protected'. Illiberal justifications, on my terminology, are those that seek to impose some notion of the good that is derived independently of the individual's own goals, values, desires and other important conative states. The classic illiberal justification for paternalism is to simply say something like 'she will be happier' or 'she will be healthier' or 'she will live longer', with no consideration of how the person's happiness, health or longevity fits into her own goals and values.

The cases that I discuss in this paper are stock examples for those arguing against the liberal approach to paternalism (e.g. Culver and Gert 1990). The case for intervention is obvious, so the reasoning goes, and yet it cannot be explained as the product of mental incompetence. The conclusion, they suggest, is that the core of the liberal approach is false: not only can we discern that the fundamental values motivating the patient's decision-making are *objectively* unreasonable, but we are morally bound to impose our more reasonable values onto the mentally competent patient (Culver and Gert 1990). Whilst that point requires a more comprehensive rebuttal than I can provide here, it is not an approach that I intend to pursue. If it is the objective unreasonableness of the patient's *values*, rather than the process by which they are formed, that justifies our paternalistic intervention, then we have no reason to single out groups on the basis of impediments such as mental illness and addiction. If we are justified in responding to changes wrought by mental illness, indoctrination or addiction with paternalism that we wouldn't impose upon mentally healthy people who hold the same values, then the appeal of such paternalism must come from the manner in which these mental impediments have altered the subjects' decision-making processes.

I should point out that whilst I raise mood disorders, addictions and indoctrination as examples, I emphasise again that I am only addressing a narrow subset of these conditions. Very often, people bearing such conditions appear to suffer inner conflicts that would make their choices less autonomous and less reasonable. For example, we could plausibly suggest that the addict wants to quit and experiences her addiction as a constraint upon her will (Frankfurt 1988, 11–25), or that the mood disordered patient's suicidal thoughts are at odds with his more settled goals (Feinberg 1984, 106–117), or that she would reject the thoughts if she were to calmly reflect upon them (Feinberg 1984, 106–117; Kleinig 1983, 67). If so, we might justify paternalistic intervention on the ground that these choices reflect momentary weaknesses, less important to these persons' autonomy than the more central projects that we are trying to protect (Kleinig 1983, 63).

But that would just skirt around the more interesting problem. Often it is not the presence of inner conflict that makes these cases so distinct, but the sheer single-mindedness with which such people pursue goals that were, not long beforehand, completely alien to them. People subject to judgment shift, such as the mood disordered, the newly indoctrinated and the psychologically addicted, can hold their new goals with as much conviction, identification and (if pushed) reflective endorsement as anyone else, if not more so. Returning, to the Joyce Brown case for illustration: the initial judge found it convincing that Ms Brown authentically preferred homelessness to psychiatric treatment and could give coherent and reflective reasons for her choice (Failer 2002, 2–26). Those who have read the description in the footnote may doubt Failer's and the Court's findings that Ms Brown was competent. For the sake of argument, however, let us take the case at its most interesting – that the Court was indeed correct in judging her competent, or at least judging there to be a lack of

evidence of incompetency. The reason why the case is interesting for my current purpose is that the appellate court did not overturn the decision regarding Ms Brown's apparent mental competency. Instead, the appeal court emphasised the *change* that schizophrenia had imposed upon Ms Brown's preferences: that only a few years earlier she had a home, highly skilled managerial employment and stable relationships with her two sisters (Failer 2002, 2–26). Intervention was justified *despite* Ms Brown's reflective preferences and ranking of values, because of the manner in which her personality had changed. It is these cases of newfound devotion that I am most interested in, where a person develops new goals or values under questionable circumstances, rather than those marked by uncertainty and conflicting desires.

4. The other self thesis

If one accepts at least the broad thrust of the previous two sections, the 'uncontroversial' justifications of mental incompetence and conflicting goals do not cover a broad class of cases where paternalism is intuitively appealing. The source of this appeal seems to be the manner in which the relevant people have changed. Put in very basic terms, it seems to many that whilst these people are choosing consistently with the goals and values that are central to their character, their character has been altered in a manner that delegitimises their preferences, such that the resulting choices are neither reasonable nor an effective exercise of autonomy. It is in this context that I believe it is worth reconsidering a theory about personal identity that has been resoundingly rejected in both bioethics and broader moral philosophy (R. Dworkin 1993, 230; DeGrazia 2005, 79–84; Ott 2009; Bluestein 1999), i.e. that: *there is a meaningful sense of 'self' that (a) is required for autonomy and rational agency, (b) persists over time, and (c) is severed in cases of the kind discussed in sections 1 to 3 (the 'other self' thesis)*.

Rebecca Dresser (1984; 1995; 1994) drew attention to the other self thesis in bioethics, as a line of argument in favour of paternalistically overruling advance directives that seek to refuse life-extending medical treatment in the event that their author becomes severely mentally impaired. She adopts a Parfitian account of personal identity to argue that a severely mentally impaired patient is a different person, i.e. a numerically different individual, to the person who executed the advance directive. My suggestion is less radical, and avoids the widespread moral revisionism that some believe is implied by such an interpretation of Parfit (e.g. Buchanan 1988). Most importantly, I believe that neither the types of cases that I have raised, nor those discussed by Dresser, involves the cessation or creation of any individual being, i.e. no-one becomes a numerically different person or a numerically different human organism. Instead, I argue that there is a *second* persistence relation, other than our persistence as numerical individuals, that we require for moral autonomy and responsibility. That is, we require persistence *of character*.

I am not suggesting that we view character as some miniature numerical identity, a person-within-a-person, such that the mentally impaired are strangers to their former selves. Instead, we should view character as a *psychological relation tying us to our previous or future choices*. This introduces temporal persistence without implying numerical distinctiveness; it is a psychological relation between us and our specific choices, rather than between past and future selves. We may have the required persistence of character with regard to some choices, whilst lacking it for other choices made at the same time. The question of character persistence is not: 'is she the same characterisation self that she was when she made the choice', but 'does her characterisation identity persist with regard to that choice'.

Any plausible account of character needs to address the observation that change to our character, without loss of moral responsibility for decisions motivated by the altered character traits, is a routine part of our ordinary existence. Unless we intend to produce a radically revisionist account, it must be possible for our character to change and develop without loss of responsibility. I will suggest that rather than mere change to character itself, the cases discussed in sections 1 to 3 involve an inconsistency in the biological and social relations from which character is derived. As such, the people who are subject to the change can not reasonably identify with both the biological and social relations that form the character traits relevant to their decision *and* contemporaneously relate to their former selves, and are thus alienated from those selves. This is a broad topic, and my ambitions in this paper are limited: I hope merely to establish that we ought to be viewing character, as it matters to autonomy and responsibility, in terms of temporal persistence. In doing so, however, it is necessary to clarify where my views sit in relation to the narrative approach championed by Schechtman (1996, 94–126), DeGrazia (2005, 78–107) and Quante (2007; Quante 1999). These authors, in similar but certainly *not* identical terms, describe us as (using DeGrazia's version for brevity), 'the individual who is realistically depicted in your self-narrative or inner story' (DeGrazia 2005, 84). I see my position as having much in common with these views, but either extending or diverging from them on the questions of how to view radical alterations in one's story, and the importance of how such changes are caused.

5. What kind of self is required for moral persistence?

Questions of moral persistence require a more substantive conception of character than numerical identity, by itself, can provide. The persistence of our moral relations of

autonomy and responsibility implies that our choices have some *current* relevance to us, and it is not clear how this is established by my remaining the same person (or the same human organism, or subject of experiences, or so on). I am the same person who at various times has needed a haircut, been a regular smoker, been unmarried and considered studying physics, but none of these historical descriptions apply to me currently. Why then, if it were the case that I am (say) the same person who, at some former time, committed genocide, would that historical description combined with numerical identity tell us anything about my current moral relations?

The question of what criteria are required for numerical persistence (the 're-identification question') is different from that of explaining *who* we are in a qualitative and a moral sense (the 'characterisation question') (Schechtman 1996, 1–14; DeGrazia 2005, 78–83). Of course, the answers to those questions are likely to be related. The answer to the characterisation question might tell us why someone might be responsible for her actions in committing genocide 30 years ago, but this is still likely to involve the observation that it was her – the same numerical individual – that committed those actions. As I have stated, I believe that in cases of the kind I discussed at the start of this paper – which I will continue to refer to as example of 'judgement shift' – illustrate a severance of character, rather than of numerical persistence.

There is nothing particularly extraordinary about citing a person's divergence from their ordinary character as a justification for paternalism, even within liberalism. The Kantian tradition has long viewed reasonableness, often measured against a person's deep-held goals and values, as necessary for effective autonomy (Rawls 1971, 248–250; Brock 1988; Scoccia 1990). Kleinig (1983, 67–68) closed much of the practical

gap between the two traditions by recognising that the Millian concept of autonomy could only warrant the overriding value attributed to it if defined in terms of personal integrity, i.e. as protecting a person's more central goals ahead of transitory desires. Faden and Beauchamp cautiously prefer a requirement of *non-repudiation* that, whilst providing a more inclusive conception of autonomy than Kleinig's, endeavours to capture a similar intuition that autonomy requires a degree of authenticity concerning our more central goals (Faden and Beauchamp 1986, 262–269) (see endnote for a brief discussion of Faden and Beauchamp's concerns about making authenticity a pre-requisite for autonomy[vi]) Whilst they were a sharp development from other Millian accounts of that time (e.g. Arneson 1980), Kleinig's views follow from a similar view of the self to that which would later be so effectively championed by Frankfurt (1988, especially 11–25; 1999, 129–141) and G. Dworkin (1989; 1988). Kleinig argues:

> *'Our lives do not always display the cohesion and maturity of purpose that exemplifies the liberal ideal of individuality, but instead manifest a carelessness, unreflectiveness, short-sightedness, or foolishness that not only does us no credit but also represents a departure from some of our own more permanent and central commitments and dispositions'* (Kleinig 1983, 67)

This reflects the common observation that at times some people feel as though they have a greater, or more important freedom, by denying their immediate desires (or having them denied to them) in favour of more important goals. For Kleinig, the role of autonomy is to protect these 'central commitments and dispositions', and so even mental competence should not be an absolute barrier to paternalism where:

> *'our conduct or choices place our more permanent, stable and central projects*

> *in jeopardy, and where that comes to expression in this conduct or these choices manifests aspects of our personality that do not rank highly in our constellation of desires'* (Kleinig 1983, 68)

I agree almost entirely with Kleinig on this particular point. However, it neither justifies paternalism, nor explains paternalism's appeal, in the 'judgement shift' cases, or at least not that subset that I specified in section three as being the cause of my concern. Unlike the scenarios that Kleinig envisages, these are not cases where someone has succumbed to some transitory weakness at the potential cost of their more central goals. Part of what makes these subsets of non-delusional mental disorders, addictions and indoctrinations seem so bizarre (even compared to other mental disorders and addictions) is the utter conviction with which the people subject to them seem to embrace their new and destructive goal. In such cases, the newly imposed goals and values are not endangering their more central goals, but have *usurped them*, such that the newly imposed goals are now the goals most central to the person's self.

As noted in section three, I recognise that this may be atypical for such conditions, and that Kleinig's tale of weak-willed lapses and internal conflict might be more useful as a general account of addiction and some mental disorders. But it would be bizarre if an addiction or indoctrination were any less deserving of paternalistic intervention simply because its coercive power was so strong that it entirely usurped the sufferer's central goals rather than merely competing with them. The judgement shift phenomena requires an account that not only explains the relationship between autonomy and character, but which tells us why that relationship sometimes seems to break down such that paternalism appears warranted even though a person is acting upon the goals

that are most central to her character. Kleinig is entirely silent on this point. We need an account of how our character *change* relates to autonomy, and that takes us instead to the accounts of characterisation identity.

The point at where my claim diverges from the established philosophical landscape is that the most influential attempts to address the characterisation question portray our character as being unified (as does my account), but in a manner that is independent of any persistence relation other than numerical persistence (unlike my account). Schechtman (1996, 1–2), DeGrazia (2005, 60–78) and Quante (1999) have led modern discussion of the characterisation question, introducing the distinction between the re-identification and characterisation questions and building upon it an account that grounds the unity of character in the narrative quality of our experiences rather than a further metaphysical persistence relation. Whilst the explanatory usefulness of 'narrative' as a concept has been far from universally accepted (Christman 2004; Strawson 2004), the underlying view that our character is revealed in our manner of experiencing the world is widely held (e.g.Bluestein 1999; Battersby 2006; Cartwright 2006; Christman 2005; Christman and Anderson 2005; G. Dworkin 1989; Edwards 2007; Atkins 2004; Frankfurt 1988, 11–25; Frankfurt 1999, 108–141; Korsgaard 1989; Kuczewski 1999; Meyers 2005; Oshana 2005; Velleman 2005, 56–76; Roesler 2006; Strawson 2004).

The foundation of the narrative view is the intuition that personhood requires us to see ourselves in a certain kind of way, one in which, as Schechtman (1996, 96) puts it, we 'conceive of [our] life as having the form and logic of a story – more specifically the story of a person's life'. Part of the appeal of this view is that it seems to explain why our past choices have current moral significance, without having to posit some

metaphysical persistence relation other than that we are numerically unified, albeit constantly changing, subjects of experiences. As such, it explains the temporal component of our character without having to demonstrate that our character has temporal persistence. In explaining why someone is morally responsible for a decades-old genocide, we don't need to show that he still holds the same values or other psychological traits that were relevant to his decisions at that time. According to the narrative view, our identity is not derived directly from our experiences and psychological traits, but rather from our *interpretation* of those experiences, whereby we render them intelligible as part of one's life as a person and moral agent (DeGrazia 2005, 84; Schechtman 1996, 96). We do this – and can *only* do this – by interpreting those experiences through the framework of a narrative: such that each time-slice gains its intelligibility from its *context* amongst what comes before and after in the narrative, such that we understand our experiences as being episodes in a unified life. Our past choices are significant because we experience them as being significant, and we do this because it is necessary in order to give our experience of life the intelligibility that we require for personhood.

I find the emphasis on 'life as we experience it' that informs the narrative account to be highly plausible as a starting point for discussion of moral persistence. Our sense of temporally unified agency, and the intuition that other persons ought to be held accountable as temporally unified agents, is something that we bring with us to moral debate and self-analysis. It is plausible that this sense of persistence is a feature of our subjective standpoint rather than something we recognise in our objective metaphysical constitution. One might suppose that there could be some external constraint upon persistence imposed by the requirement that our experiences be amenable to narrative interpretation. However, critics of narrative account have often

pointed out that there no obvious boundaries on what events could count as a narrative, and Schechtman's (1996, 96) own account emphasises that the metaphor merely refers to temporal linearity and a need for contextual interpretation[5] rather than any particular shape. Moreover, as Schechtman (1996, 99) hints, if our lives were not amenable to narrative interpretation, we might still interpret them in a manner that is non-linear but no less unified. The interesting part, for my present purpose, is not the descriptive usefulness of the term 'narrative', but the notion that our moral persistence arises from our interpretation of our experiences. I find this to be a highly plausible starting point for discussion of moral persistence. Just not a very good end point.

I know I am not alone in this, as Quante (2007) makes much of the socially constructed aspect of character, stating that 'a human person's personality is essentially constituted by social relations' (2007, 56). Our narrative self-conception, on Quante's view, is not something we create individually, but develops within a social context. Quante is quite correct on this point, but it is not at all inconsistent to also acknowledge that individuals do not always *fully* absorb the social norms of their environment – no matter how strong the social prohibition is on killing, or on theft, there are always those occasional individuals who not only kill or steal but see nothing wrong with people doing so. Even if there weren't such people, we still might not be satisfied with a naturalistic explanation, but could want some normative (or meta-normative) justification of why we *should* abide by the social norms that require us to take responsibility for past infractions.

---

[5]. We can grant the linearity requirement as specifying the only plausible type of narrative that personhood could involve, but as a general description of narrative it is unclear whether even that stipulation is a strict requirement. Postmodern novels, such as Don Dellillo's Pulitzer Prize nominated *Underworld* (New York: Scribner, 1997), construct narratives from non-linear sets of events, endeavouring to draw interpretations (in DeLillo's novel, to the history of American popular culture) through juxtapositions that would be excluded by a linear approach.

DeGrazia (2005, 84) says '*You are the individual who is realistically depicted in your self-narrative or inner story'*. But people can be mistaken both about the kind of person they are and their relation to their past actions. Perpetrators of genocide can *erroneously* believe that they are no longer responsible for their crimes many decades ago. DeGrazia (2005, 85) and Schechtman (1996, 114–122) exclude or amend personal narratives that are factually delusional, such as a claim that one was turned into a snake, or a narrative that is incapable of articulation, but such constraints do not address the more common difficulty of mentally *healthy* persons who misinterpret their experiences to deny moral persistence. Their narratives are not flawed through factual error but through erroneous *interpretation* of the facts, and to exclude accounts on that ground would be to throw out Schechtman's and DeGrazia's central claim that self-conception is what matters.

This is less troubling if, like Schechtman (1996, 95–97), we adopt the narrative account as an explanation of what is necessary for healthy personhood, rather than a blunt statement about our self-conception. The narrative nature of our experiences may be what allows us to impose that unity, but we need such unity because personhood requires that we see ourselves as active participants in our own development, capable of agency and self-authorship rather than as passive heirs to our prior causes. Philosophy has long been fixated with the tension between this requirement and the observation that we cannot logically escape the biological and social relations that comprise and influence us. The Frankfurt/Dworkin tradition, which the narrative account of character appears to have grown from, developed partially in response to this tension and their answer has been that our experience of agency and autonomy involves our *identifying with* the influences that our biological and cultural relations

impose (Frankfurt 1988, 11–25; Frankfurt 1999, 129–141; G. Dworkin 1989; Wolf 1989; Oshana 2005; Meyers 2005; Christman 2004; Waldron 2005, 307–329). What matters to autonomy is not that we exercise ultimate control over the desires and values that guide our choices, but that those which guide us are those that we identify with as comprising our will rather than those that we experience as unwelcome and external constraints.

But what happens when those biological and social relations change so rapidly and drastically, and from such a divergent source, that we cannot plausibly identify with both their influence and our pre-change lives? That is what seems to happen in cases of judgment shift, biologically in the case of mental illness and addiction, and culturally with regard to indoctrination. In order to retain personhood, the people involved *must* identify with the current biological and social relations that constitute them (though not in so explicitly self-conscious a manner), yet they cannot reasonably do this while also identifying with their past self.

This is most easily illustrated with regard to mental illness. It is a conceptually vital feature of illness that it is an external constraint upon us rather than an aspect of our character (Boorse 1975; Wakefield 1992; Nordenfelt 2007; Fulford 2001; C. Edwards 2009). People with illness (mental or physical) do not thereby have bad character, and those fortunate enough to be particularly resilient to illness do not thereby have good character. Working within the same social and cultural framework identified by Quante, we have drawn a normative distinction between the brain processes comprising mental illness and those that comprise healthy brain function. We cannot, then, identify mental illness as being part of the same unified self that is comprised by our healthy biological and social relations. There is no problem inherent in

experiencing personality change that has been inspired by something external to oneself. However, the changes induced by mental illness that concern us are not 'inspired by' mental illness in the same way that we may be inspired to character change through a friendship or by migrating. Instead the changes are part of the mental illness itself; the illness is comprised by changes in the way that our brain processes operate, and so by becoming mentally ill we externalise part of those brain processes.

In raising this, I hope to have drawn attention to a by-product of our viewing selves as having temporal narrative status: that the burden required for healthy personhood has been raised. The narrative account gives us reason to view personhood as requiring a temporal unity that is not always achievable. Western culture is accustomed to defining self and personhood so as to exclude the brain processes that comprise mental illness, addiction and indoctrination by drawing a distinction between the biological and social relations that comprise those conditions and those which comprise healthy personhood. We cannot be *both* a person formulated through such norms *and* one that radically rejects them by identifying with brain processes that our cultural context would ordinarily view as disordered. Such a change would require that we view ourself as having adopted a new identity, severed from our prior self-narrative; effectively, a different person. The phenomenon of becoming mentally disordered involves our having adopted brain processes that, through our cultural norms or personal narrative, we cannot plausibly unify with our prior personality. The narrative approach leads us to viewing these brain processes as something external to one's self, part of what one's healthy narrative is defined against. That is why paternalism is so appealing in these cases; we are not seeking to protect a person's wellbeing by imposing on his liberty, but are defending his own goals and values against something we see as external and incompatible with his genuine self.

6. In conclusion: the sources of character development

What this illustrates is that the *sources* of our character development matter. Mental illness can sever our character by forcing a conflict between the biological relations that currently comprise us and those that comprise our former selves. Our distinction between illness and self gives us reason to distinguish between the brain processes that result from mental illness and the modes of thinking that we employ when mentally healthy. The character change involved then, is not part of our ordinary growth and development, but arises from a change in our composition that is not amenable to the narrative unity that ordinarily comprises our personhood. We could make similar claims regarding addiction and indoctrination, that again are made possible by the conceptual distinction between the biological and social relations that comprise those alterations and those which comprise healthy personhood.

The narrative account rests on the presumption that we maintain the biological and social relations that make our ordinary state of unified personhood possible. The cases that I have described as judgment shift involve change to those relations, such that they are conceptually distinct from those which previously constituted us, as in the distinction between mental illness and healthy brain function. I describe this as a persistence relation because it is the *change* – the incompatibility of the pre-shift and post-shift states – that causes the severance. The temporal incompatibility involved in judgment shift becomes relevant because cases remain where mental illness will severely harm a person contrary to his views when healthy, but in circumstances where we cannot readily argue that the person is incapable of the reasoning and insight required for mental competence. On the presumption that we have good reason to prefer the earlier self's goals to that of the latter, i.e. that a self constituted by healthy

biological and social relations has greater moral legitimacy than one constituted by 'ill' biological and social relations, the common intuition that paternalism is warranted because a person's behaviour is vastly 'out of character' need not be overridden by our inability to demonstrate that the person is mentally incompetent.

The emphasis upon mental competence was adopted largely from well-grounded liberal fears of wrongly imposing our goals and values upon others, resulting in our preference for judging a person's ability to reason regarding her goals rather than making any judgments about what her goals should be. The temporal approach to character gives us an alternative that does not involve imposing any values but the person's own. In such cases, we intervene to protect the person's *own* goals and values from those which have been wrongfully imposed by sources that are external to that person. What better reason for intervention could a liberal desire?

References

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual 4th edition: Text Revision*, American Psychiatric Publishing.

Anderson, Joel, and Axel Honneth. 2005. 'Autonomy, Vulnerability, Regonition, and Justice'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 1–23. Cambridge: Cambridge University Press.

Archard, David. 2008. 'Informed Consent: Autonomy and Self-ownership'. *Journal of Applied Philosophy* 25 (1): 19–34.

Arneson, Richard. 1980. 'Mill Versus Paternalism'. *Ethics* 90 (4): 470–489.

Atkins, Kim. 2004. 'Narrative Identity, Practical Identity and Ethical Subjectivity'. *Continental Philosophy Review* 37 (3): 341–366.

Battersby, James. 2006. 'Narrativity, Self and Self-representation'. *Narrative* 14 (1): 27–44.

Bluestein, Jeffrey. 1999. 'Identity Revisited'. *Journal of Law, Medicine and Ethics* 27: 20–31.

Boddington, Paula. 1998. 'Organ Donation After Death - Should I Decide, or Should My Family?' *Journal of Applied Philosophy* 13 (1): 69–81.

Brock, Dan. 1988. 'Paternalism and Autonomy'. *Ethics* 98 (3): 550–565.

Buchanan, Allen. 1988. 'Advance Directives and the Personal Identity Problem'. *Philosophy and Public Affairs* 17 (4): 277–302.

Cartwright, Will. 2006. 'Reasons and Selves: Two Accounts of Responsibility in Theory and Practice'. *Philosophy, Psychiatry, and Psychology* 13 (2): 143–155.

Christman, John. 2004. 'Narrative Unity as a Condition of Personhood'. *Metaphilosophy* 35 (5): 695–713.

———. 2005. 'Autonomy, Self-Knowledge and Liberal Legitimacy'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 330–358.

Cambridge: Cambridge University Press.

Christman, John, and Joel Anderson. 2005. 'Introduction'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 1–23. Cambridge: Cambridge University Press.

Culver, Charles, and Bernard Gert. 1990. 'The Inadequacy of Incompetence'. *The Millbank Quarterly* 68 (4): 619–643.

Davis, John. 2008. 'How to Justify Enforcing a Ulysses Contract When Ulysses Is Competent to Refuse'. *Kennedy Institute of Ethics Journal* 18 (1): 87–106. doi:10.1353/ken.0.0001.

DeGrazia. 2005. *Human Identity and Bioethics*. New York: Cambridge University Press.

Dresser, Rebecca. 1984. 'Bound to Treatment: The Ulysees Contract'. *The Hastings Centre Report* 14 (3): 13–16.

———. 1994. 'Missing Persons: Legal Perceptions of Incompetent Patients'. *Rutgers Law Review* 46 (2): 609–719.

———. 1995. 'Dworkin on Dementia: Elegant Theory, Questionable Policy'. *Hastings Centre Report* 25 (6): 32–38.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

———. 1989. 'The Concept of Autonomy'. In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman. New York: Oxford University Press.

Dworkin, Ronald. 1993. *Life's Dominion*. New York: Knopf.

Edwards, Steven. 2007. 'Disablement and Personal Identity'. *Medicine, Health Care and Philosophy* 10: 209–215.

Faden, Ruth, and Tom Beauchamp. 1986. *A History and Theory of Informed Consent*. New York: Oxford University Press.

Failer, Judith. 2002. *Who Qualifies for Rights?: Homelessness, Mental Illness, and Civil Commitment*. Ithica: Cornell University Press.

Feinberg, Joel. 1984. *Harm to Self*. Vol. 3. 3 vols. The Moral Limits of the Criminal Law. New York: Oxford University Press.

Frankfurt, Harry. 1988. *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press.

———. 1999. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Friedman, Marilyn. 2005. 'Autonomy and Male Dominance'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 150–173. Cambridge: Cambridge University Press.

Kleinig, John. 1983. *Paternalism*. Totowa: Rowan and Allenheld.

Korsgaard, Christine. 1989. 'Personal Identity and the Unity of Agency: A Kantian Response to Parfit'. *Philosophy and Public Affairs* 18 (2): 101–132.

Kuczewski, Mark. 1999. 'Commentary: Narrative Views of Personal Identity and Substituted Judgement in Surrogate Decision Making'. *Journal of Law, Medicine and Ethics* 27 (1): 32–36.

Meyers, Diana. 2005. 'Five Faces of Selfhood'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 27–55. Cambridge: Cambridge University Press.

Mill, John Stuart. 1977. *On Liberty.* Ed. John Robson. The Collected Works of John Stuart Mill. Toronto: University of Toronto Press. http://oll.libertyfund.org/title/233.

Oshana, Marina. 2005. 'Autonomy and Self-Identity'. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 77–97. Cambridge: Cambridge University Press.

Ott, Andrea. 2009. 'Personal Identity and the Moral Authority of Advance Directives'. *The Pluralist* 4 (3): 38–54.

Quante, Michael. 1999. ‘Precedent Autonomy and Personal Identity’. *Kennedy Institute of Ethics Journal* 9 (4): 365–381.

———. 2007. ‘The Social Nature of Responsibility’. *Journal of Consciousness Studies* 14 (5-6): 56–76.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Roesler, Christian. 2006. ‘A Narratological Methadology for Identifying Archetypal Story Patterns in Autobiographical Narratives’. *Journal of Analytical Psychology* 51: 574–586.

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornwell University Press.

Scoccia, Danny. 1990. ‘Paternalism and Respect for Autonomy’. *Ethics* 100 (2): 318–334.

Strawson, Galen. 2004. ‘Against Narrativity’. *Ratio* 17: 428–452.

Velleman, David. 2005. ‘The Self as Narrator’. In *Autonomy and the Challenges to Liberalism*, ed. John Christman and Joel Anderson, 56–76. Cambridge: Cambridge University Press.

Waldron, Jeremy. 2005. ‘Moral Autonomy and Personal Autonomy’. In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman and Joel Anderson, 307–329. Cambridge: Cambridge University Press.

Wolf, Susan. 1989. ‘Sanity and the Metaphysics of Responsibility’. In *The Inner Citadel: Essays on Individual Autonomy*, ed. John Christman, 137–151. New York: Oxford University Press.

End Notes

[v]. On 28 October 1987, Ms Brown was examined and detained for emergency psychiatric treatment by members of New York City's Homeless Emergency Liaison Project (Project HELP). Workers from Project HELP had been monitoring Ms Brown from December 1986, and Ms Brown had been living on the street since at that time at the latest. During the interim, they had observed Ms Brown engaging in bizarre behaviour, including eating money, tearing money up and urinating on it, throwing money on the street, speaking to herself with abrupt changes of affect as if in response to some unseen stimuli, gesturing as if to a conversant who wasn't present, expressing paranoid delusions about black men trying to turn her into a prostitute (allegedly her reason for destroying money), baring her buttocks, wandering into traffic without regard for her safety, and chasing imaginary people. Her clothes smelled of faeces and appeared tattered, dirty and inadequate for living outdoors in the cold New York winter. Following her detention, the psychiatrist Lincoln Hess of Bellevue Hospital applied for Court orders for involuntary hospitalisation and treatment of Ms Brown. Evidence was presented that prior to developing mental illness, Ms Brown had held responsible positions at Bell Laboratories and a Human Rights Commission, and had regular contact with family.

At hearing, there was conflicting psychiatric evidence both as to the nature of Ms Brown's mental illness and the extent of its effects, but two features stand out. Firstly, Ms Brown was mentally competent insofar as she was able to understand her actions, explain her reasons for them and comprehend and analyse the likely repercussions of those actions. Secondly, Ms Brown had experienced a horrific deterioration in her quality of life, seemingly pursuant to a drastic change in her character. In short, Ms Brown was capable of effectively expressing her autonomy by reasoning about her goals and desires and putting them into action, but those goals had become seemingly bizarre, especially compared to those of her former self.

At the first hearing, Justice Lippmann refused the application for involuntary treatment, ruling that Ms Brown's mental competence gave her the right to choose her lifestyle no matter how bizarre her priorities. The Appellate Division overturned this decision by 3-2 majority. The majority judges focussed upon the deterioration in Ms Brown's lifestyle, reasoning that the appropriateness of involuntary hospitalisation should be considered in view of this change in character and lifestyle, rather than just Ms Brown's immediate mental competence. Ms Brown then appealed to the New York State's highest court, but the appeal was never heard. While the case was on appeal, the Hospital applied to the Supreme Court for permission to involuntarily medicate Ms Brown (the previous order only applied to hospitalisation, not medication). After hearing evidence from a Court-appointed psychiatrist, the Supreme Court ruled that Ms Brown was at least partially mentally competent to make decisions regarding medication, and refused the order, this time apparently focussing upon Ms Brown's *current* decision-making capacity. Being unable to medicate Ms Brown, the hospital discharged her, following which the appeals court ruled that the appeal was now moot and would be disposed of without a judgment (Failer 2002).

[vi] Faden and Beauchamp are concerned that a requirement of reflective endorsement, and perhaps even (to a lesser extent) their preferred requirement of non-repudiation, may allow unjust intervention in decisions that are made deliberately and with much informed consideration, due to the person failing to reflect upon some of the moral and

philosophical principles that underly her decision. In particular, they cite Miller's willingness (in Miller, 1981 'Autonomy and the Refusal of Life Saving Treatment' *Hastings Center Report*, 11: 22) to claim that doctors should assist patients to reconsider inauthentic decisions and assist them in effectively deliberating and reaching authentic treatment choices, as evidence of a risk that an authenticity requirement may encourage physicians to unjustly override patient decision-making. It is difficult to understand this claim as being anything but a slippery slope fallacy. The threat of moral hazard is only present if doctors were to override, rather than assist, patient decision-making. The insistence that doctors assist authentic decision-making, rather than simply overriding inauthentic decisions, is merely an example of the simple but widely accepted rule that paternalism should always take the least intrusive form that is sufficient to avoid harm that hasn't been autonomously chosen by the sufferer. The concern that minimal intervention could lead to unjust intervention could be applied to almost any conceivable example of paternalism, and any intervention must be premised on the grounds that intervention can be restricted to just limits. It is true that many decisions which we would ordinarily describe as autonomous do not involve reflective endorsement. But in almost all such cases, once the most minimal intervention (i.e. mentioning that those factors might be worth considering) has been tried, the person will then have reflectively endorsed or rejected her decision
Faden and Beauchamp also appear to overlook a vital feature of the authenticity requirement: that its aim is not to impose some objectively 'best' outcome, but rather to ensure that the patient's *own* central goals and desires are protected. It is unclear why Faden and Beauchamp believe that allowing a patient to unreflectively jeapordise their own more central goals is any more liberal or autonomous than paternalistically protecting those goals unless the patient reflectively decides otherwise. It is difficult to see what better justification for paternalistic intervention there could be than to protect the things that the patient himself cares for the most.

Paper 6 - Respect for other selves

Abstract: Philosophers have mostly advocated that advance directives should bear the same authority, with regard to refusal of life-extending treatment, as a patient's contemporaneous consent or refusal. Such authors typically support this position through a theory of persistent personal identity. I agree that the loss of mental competence does not render someone a moral stranger to their prior goal, but argue that equating advance direction with consent is to ignore the capacity of non-persons to attribute and withhold moral value. A distinction should be drawn between advance directives that seek to pursue deeply held goals, and those that express contempt for the mentally incompetent. While this article deals specifically with advance directives, and the conflict between the instruction to refuse life-extending treatment and the interests of the current patient, the interests of the pre-impairment person is not wholly dependent on its being legally recorded and the same conflict may arise even in the absence of advance directives. In addition to the concern at the heart of this paper, I address the conflict between two entirely legitimate sources of moral interest: that of a person to have her final days or years conducted in a manner consistent with her authentic goals and values, and that of a mentally impaired patient's powerful experiential interests.

Respect for other selves

1. Advance directives and the other self debate

How ought we respond to advance directives that appear to fly in the face of a severely mentally impaired patient's quality of life? An advance directive is a legal instrument wherein a person records instructions regarding the medical treatment that she is to receive in the event that she becomes persistently incapable of giving informed consent to or refusing treatment. Where these instructions are legally binding, they enable a person to exercise control over her future medical treatment. This has been welcomed by some on the grounds that it increases patient autonomy, but there has also been concern that in cases in which a patient is left conscious but severely mentally impaired, the person's advance instructions may be at odds with her future interests.

Two sources of this concern are discussed in the philosophical and bioethical literature. One is misinformation or false expectations by the advance directive's author regarding her future experience of illness and the treatments available (e.g., Dresser 1995). The other, which is the subject of this paper, is the sacrifice of such interests as the promotion of future happiness or the prevention of suffering by the advance directive's author in the pursuit of nonexperiential goals and values. The most obvious example of this is the use of an advance directive to refuse life-extending treatment irrespective of whether the patient is happy and content following the onset of advanced dementia. In making the advance directive, the person is not seeking to safeguard her future happiness but rather her current goals such as (her conception of) dignity or independence. The apparent tension between the advance directive and the patient's interests, then, is twofold: not only is the advance directive at odds with the promotion of the patient's ongoing happiness but it is motivated by goals and values of

a kind that the patient, in her severely impaired state, can no longer evaluate and no longer cares for.

For philosophers, advance directives exemplify a kind of concern where practical questions of practice and policy meet long-standing conflicts of philosophical principle. The authority of patient instructions in medicine is usually determined by the doctrine of informed consent, in which the patient's instructions are judged not by their reasonableness but by the patient's mental competence and access to information (Kirby 1983; Kihlbom 2008). However, advance directives take effect at a time when the patient is no longer capable of giving informed consent. The doctrine of informed consent is the high watermark of philosophical and ethical liberalism, in which personal autonomy is given an overriding authority that is increasingly rare in other parts of life. Conflict between advance directives and patients' interests invites a consideration of how the loss of personhood affects the claim to personal autonomy and integrity on which philosophical liberalism (and hence the informed consent doctrine) rests.

Ronald Dworkin discusses the problem in *Life's Dominion* (1993, 220–30), where he recounts the experience of a medical student who repeatedly visited "Margo," who was then suffering the advanced stages of Alzheimer's disease. Margo's mental capacities are impaired to the point that she is only capable of a daily routine of simple activities, but she appears to obtain substantial pleasure from those activities. The student describes her as possibly one of the happiest people he has ever met. Dworkin then asks, counterfactually, what our appropriate response should be if Margo had (while mentally competent) executed an advance directive requesting that any life-extending treatment be withheld once she reached her current advanced stage of dementia.

We can describe Margo as having reached a "post-personhood" stage in her life. "Personhood" is often used loosely to refer to the mental capacities that we associate with the typical adult human. We can also use the term in a stricter sense, to refer to the property of holding the capacity for personal autonomy and moral responsibility. Personhood, in this latter sense, can be decision-specific: in an earlier stage of dementia, Margo may have been capable of personal autonomy in relation to medical procedures while lacking that capacity with regard to the distribution of her estate. However, now Margo has lost personhood in a broader sense. Advanced dementia has eroded the "self" on which Margo's personal autonomy and moral responsibility depends. That is, it has deprived her of the capacity for long-term memory, forethought, planning and reflection that could enable a sense of persistent identity and a conception of the kind of life she wants to lead. Margo's state is one of "post-personhood" in that she no longer comprises a sophisticated self to which qualities like personal autonomy or moral responsibility can apply.

Margo's case is troubling because it reveals a conflict between two sets of interests arising from different periods in Margo's life, *both* of which we would normally view as obliging our respect. The practice of advance directives is grounded in our moral claims to autonomy and the pursuit of our deeply held goals and values, these being interests that arise from our existence as *persons*. By directing that her death be hastened in the event of severe mental impairment, Margo brings these autonomy interests into conflict with the persistence of her own post-personhood life. However, Margo's post-personhood life is characterized by happiness and contentment—experiences that imbue her life with value. The interests underlying the moral authority of Margo's advance directive—namely, her interests in autonomy and the pursuit of

her preimpairment goals and values—are in conflict with her interest in continuing to live a severely mentally impaired but happy life.

There is a sense in which this is a conflict between different stages of Margo's life. When making her advance directive, Margo (assuming she was well-informed and instrumentally reasonable) placed greater worth on the goals and principles motivating her refusal of treatment than she did on the prospect of a happy post-personhood life. After the onset of advanced dementia, however, Margo is no longer capable of having concern for complex goals and principles but nevertheless remains capable of experiencing happiness and contentment. This is troubling for those of us who believe that Margo's "own" wishes should be applied; we would need to ascertain whether to prioritize the way of living that Margo *currently* values or to impose the instructions she gave prior to losing the capacity for personhood.

This conflict has encouraged philosophers to view Margo's different stages as different "selves." Under this view, "Margo, the person," who gives rise to Margo's interests in autonomy and the pursuit of goals and values, forms one relevant self, whilst "Margo, the severely mentally impaired human," who has ongoing experiential interests, is another such self. This terminology does not necessarily imply that the selves are numerically distinct individuals but rather simply marks a disunity in Margo's character (i.e., in the conative states—goals, values, attitudes, and principles—that characterize Margo's life). Dworkin reasons that we should consider Margo's current state to be an extension of "Margo, the person," and that respect for both Margo's autonomy and her best interests should lead us to implement the advance directive (1993, 231). Other authors have thought differently. Rebecca Dresser (1994; 1995), Jeff McMahan (2002) and John Robertson (1991) claim that when an individual loses

personhood (in the broad sense experienced by Margo), they no longer have a substantial prudential or moral commitment to the goals and values that characterized them as persons:

> "*An individual's good may change along with his nature, character, values and preferences. And this is the basis for claiming that it would be better for the Demented Patient now to continue to live. The dementia has caused a profound alteration in her nature. She was once devoted to creative intellectual endeavors. For a person like that, the plunge into dementia could be deeply degrading. Yet now that the dementia has already occurred, the elements of her nature that opposed or were hostile to a state of contented dementia have been eradicated. Now that she is demented, the good that seems appropriate to her present nature is contentment."* (McMahan 2002, 500)

When our situations change drastically, our interests and preferences also change. The difference between competent and incompetent interests is so great that if we are to respect incompetent persons, we should focus on their needs and interests as they now exist and not view them as retaining interests and values that, because of their incompetency, no longer apply (Robertson 1991, 7).

From this they conclude that, where a post-personhood patient's experiences and preferences are mostly positive, we should not impose any earlier directive that would deprive her of those experiences and preferences, as the interests that motivated her to make the advance directive are no longer relevant. For ease of reference, I refer to these arguments as the "other self thesis": they emphasise the *change* that has occurred since the advance directive was made, such that the patient's current experiential

interests ought to constrain our implementation of the choices made by the person she once was.

Many people find the other self thesis intuitively compelling, as is suggested by the fact that Dworkin keeps open the possibility that moral considerations external to the patient's interests may justify limitations on the use of advance directives (1993, 228–29). Nonetheless, the other self thesis has found very little philosophical acceptance, particularly in the field of bioethics. Numerous authors have disputed the claim that loss of the *capacities* that comprise personhood divests the patient of the *moral interests* arising from personhood (e.g., Blustein 1999; Kuczewski 1999; Ott 2009; DeGrazia 2005). These arguments vary in the ways that they conceive of personal identity, character, and the interests that motivate the advance directive, but they all lead to the principle that we ought to consider the patient's life as a whole, wherein the claims of the person that she was outweigh those of the human that she is.

I suspect that these arguments are correct in their common claim about personal identity: that we ought to consider Margo's life "as a whole." Whether we characterize them as separate but closely related identities or as different stages of the one identity, "Margo, the person," and "Margo, the post-personhood human," are not moral strangers. However, they are mistaken in interpreting the "other self thesis" as purely a problem of personal identity. We can acknowledge the continuing relevance of the interests arising from Margo's personhood, such as the goals and values that informed her advance directive, without accepting that those interests are the sole determinant of Margo's moral worth. If we are to care for her life "as a whole," we ought to recognize that Margo is not *just* an extension of the person she once was. Moreover, the liberal multifaceted conception of the good that underlies the moral authority of advance

directives also obliges us to recognize that Margo's way of life can have value regardless of the rejection of that way of life implicit in her advance directive. Moral worth is not restricted to personhood, and so we are obliged to respect not only "Margo, the person," but also those interests arising from Margo's post-personhood life.

In endeavoring to rehabilitate the other self thesis, I do not intend to make out a case against the advance direction of euthanasia in general. To the contrary, I hope to provide an account that allows for the advance refusal of life-extending treatment such that failure to apply an advance directive would catastrophically undermine a patient's preimpairment life projects. My aim is to show that our obligation to apply a patient's advance directive is constrained by that patient's ongoing moral worth. If such advance directives are to be of use, we must reenvisage the form that advance directives should take and the process by which they operate. We must look beyond the face of the patient's bare instructions to consider the values and goals that the advance directive promotes and the role that those values and goals play in the patient's overall life. Rather than holding authority by virtue of being a properly executed document, then, advance directives refusing treatment should be subject to the discretion of a substituted decision maker—preferably one appointed by the patient when making the advance directive—tasked with weighing the patient's preimpairment instructions against the value of her current way of life.

2. Patient Control of Medical Decisions and the Appeal to Personal Integrity

It is worth distinguishing two contexts in which we might analyze the moral significance of advance directives. Dan Brock states:

> *"Executing an advance directive is more than simply expressing a preference regarding future treatment. Advance directives, like wills governing property, make use of a rule-governed social practice in order to create and secure from others obligations and commitments about one's medical treatment in the case of future incompetence. They constitute what philosophers have called performative utterances that call on background social practices in order to create obligations and responsibilities."* (1988a, 251-252)

Brock is correct on this point: the very existence of social norms alters our moral obligations, and if a particular social norm is relied on enough, it may well become self-justifying. When asking whether an individual advance directive should be implemented, then, we need to take into account the legal and social commitments that were secured by executing it. However, in many countries, advance directives are a relatively recent practice (Bogdanoski 2009), and formal policy can be expected to guide social expectations. My aim here is to analyze advance directives as a practice, asking what obligations and commitments *should* be created by recording one's instructions regarding future treatment.

In the context of its background social practices, including the informed consent doctrine, an advance directive is a means of giving or withholding informed consent for future treatment. However, in evaluating the moral authority of advance directives as a practice, we ought to ask what moral claims underlie the authority that medical ethics attributes to a patient's informed decision to give or refuse consent, and then ask whether those moral claims are still served when we give informed consent by means of an advance directive. Of these moral claims, some can be characterized as relating

to the general value of personal *well-being*. In a medical context, "well-being" suggests physical health, but I do not want to limit its scope to physical health; rather I want to use to refer to all of a person's *experiential* self-oriented interests. Through the informed consent doctrine and the practice of advance directives, we can promote patients' well-being by promoting their happiness, improving patient-doctor relationships, and facilitating the effective provision of future medical treatment. But there are some occasions when we have good reason to believe that we can improve the well-being of an informed and mentally competent patient by imposing treatment against her will; such as where a patient refuses life-saving medical treatment on moral or ideological grounds. Nonetheless, the informed consent doctrine serves to prevent unwanted treatment in such cases, on the understanding that the person has nonexperiential interests in *personal integrity* that warrant protection from paternalism.[6] The liberal advocates of advance directives claim that this appeal to personal integrity gives rise to an interest in having one's advance directive implemented, that is similarly independent of well-being.

The appeal to personal integrity has its roots in John Stuart Mill and was adopted into modern theories of ethics as the starting point of liberal arguments against paternalism (Arneson 1980; Feinberg 1984; Kleinig 1983; Mill 1977). This statement by Ronald Dworkin is representative:

[6] This is often referred to as "personal autonomy"—I prefer "personal integrity," because while both terms have multiple meanings, the philosophical over-use of "personal autonomy" invites confusion between the numerous different but related conceptions that it can refer to (Dworkin 1988, 5–10).

> *"Recognizing an individual right of autonomy makes self-creation possible. It allows each of us to be responsible for shaping our lives according to our own coherent or incoherent—but, in any case, distinctive—personality. It allows us to lead our own lives rather than be led along them, so that each of us can be, to the extent a scheme of rights can make this possible, what we have made of ourselves."* (Dworkin 1993, 242)

The central claim is that the "good life" for a person is (among other things) one that is consistent with the person's deeply held goals, values, and convictions—that is, a life of personal integrity (Arneson 1980; Feinberg 1984, 52–97; VanDeVeer 1986, 124–27; Kleinig 1983, 27–30; Brock 1988b; Scoccia 1990). Dworkin actually makes two arguments toward this claim, one grounded in *autonomy* (1993, 224), and the other in *beneficence* (1993, 231), but both lead to the same moral prioritization of deeply held goals and values ahead of well-being. In broader moral philosophy this "central claim" requires defending, but in the subfield of bioethics (and especially in discussion of medical paternalism and informed consent) it has become part of the accepted theoretical background (e.g., Buchanan 1988; DeGrazia 2005; Harvey 2006; Kuczewski 1994, 1999; Kuhse 1999; Post 1995). I compromise by omitting the lengthy reasoning required to justify the appeal to personal integrity while explaining its implications for medical decision making.

Personal integrity is not simply a matter of pursuing one's immediate preferences. As Dworkin's passage indicates, personal integrity is a matter of design, a shaping of one's life, that stands quite at odds with an impulsive focus on transitory desires. That is why I preface my reference to the conative states (goals, values, convictions, etc.) relevant to personal integrity by describing them as "deeply held." What I mean by this

is that they accurately reflect the person's own attribution of value: the way that a person values things and experiences within her life, thus imbuing with value both the thing itself and her life insofar as it contains that thing. For example, by deeply valuing the practice of playing the piano, I give that practice value within my life and also imbue my own life with value—my life has greater worth (to me) than it would without such attribution of value.

The term "deeply held" is deliberately vague, because it refers to a concept that is easily understood at a loose level but that invites interminable difficulties when analyzed with greater precision. We all understand what it means to say that certain of our values matter greatly to us, that those values are what define our character, and that other preferences are less important to us. However, when trying to state what it is that makes a particular value more "central" or "deeper" than others, philosophers struggle to agree on seemingly basic notions such as the relevance that a value's stability, longevity, content, and formation process has over that value's importance (Frankfurt 1988, 11–25; Frankfurt 1999, 129–41; Christman 2004, 2005; Dworkin 1989; Meyers 2005; Taylor 2003; Varelius 2006; Young 1989). We can say with less controversy that the more important values are those that we identify with most strongly, whereas we do not identify with the preferences that we temporarily succumb to when our will is weak (Frankfurt 1988, 11–25; Frankfurt 1999, 129–41; Dworkin 1989; Christman 2005). However, while that may help relate the idea of "deeply holding" a value to our concepts of character and identity, it doesn't render the attribution of value itself any less opaque. For the current discussion we can use the loose notion of "deeply holding" goals and values while putting aside how the question of how that notion should be filled in.

Severe mental impairment can threaten our personal integrity by rendering us unable to effectively pursue our goals and implement our values. Our most deeply held goals and values are often those that characterise our conception of how we ought to live. Our commitment to some of these goals and values, such as our moral codes, extends to circumstances where we lack the capacity to experience or understand the effects of breaching them. For example, many of us would be horrified at the idea of being kept alive by means of an organ transplant using an organ that was stolen from a child who was killed for that purpose. The knowledge that our faltering mental state may prevent us from understanding how the organ was procured would not lessen our disgust. Some such commitments may be more esoteric—for example, a Jehovah's Witness's refusal of blood transfusions or a musician's repugnance at the idea of living without the capacity to perform Beethoven. In such cases, the person holds the goal or value to be more important than the extension of her life, even if that life is characterized by happiness or contentment.

In some cases, the patient's commitment to her deeply held goals and values can only be met by withholding life-extending medical treatment. Ordinarily, in the event of the patient's mental incompetence, a paternalistic substitution of the patient's consent would allow that treatment. Given such a scenario, then meaningful pursuit of the patient's goals and values in the face of mental incompetence requires some means of making medical decisions in advance, *before* she suffers mental impairment. If such measures are not in place prior to the patient losing mental competence in relation to her treatment choices, she risks becoming subject to a substituted decision-making procedure that is unmotivated to prioritize her goals and values ahead of her experiential interests.

So, while individual advance directives may be made for any number of reasons, the *social and legal practice of advance directives* primarily serves to protect personal integrity, with a secondary role of promoting well-being. Where a patient has been deprived of consciousness or temporarily rendered mentally incompetent, the appeal to personal integrity explains why the patient's former instructions through her advance directive remain morally authoritative. The other self thesis, however, questions whether the appeal to personal integrity has the same relevance when applied to a patient who is conscious but has permanently lost the capacity for personhood. The notion that this drastically changes the patient's interests appears compelling because, ordinarily, personal integrity concerns the goals and values that we *currently* deeply hold and identify with. There is something extraordinarily inauthentic about allowing your life to be governed according to a set of goals and values that you no longer identify with. This places the onus on opponents of the other self thesis to explain why the appeal to personal integrity continues to have relevance in such circumstances.

3. Surviving Interests

The accounts that I discuss in this section vary considerably in their conceptions of personal identity and character, but they converge through their shared proposition that whatever experiential interests Margo (I'll continue to use her as a representative case) has qua her existence as "Margo, the post-personhood human," she is also the biological extension of "Margo, the person," and therefore the beneficiary of that person's interests. My interest is not so much in the intricacies of their varying models of personal identity as in the implications of their claim that the erosion of an individual's goals and values caused by advanced dementia does not sever her interest in pursuing those former goals and values. In the later sections of this paper, I

acknowledge that this claim is compelling and argue that the other self thesis can be reconstructed in a form that is consistent with that position. I refer to these accounts collectively as the "surviving interest thesis," after their shared conclusion that the interests arising from personhood survive the loss of the capacity for personhood.

The surviving interests thesis begins with Dworkin: "When we consider how the fate of a demented person can affect the character of his life, we consider the patient's whole life, not just its sad final stages, and we consider his future in terms of how it affects the character of the whole" (1993, 230). Dworkin's view invites the response that he is asking us to consider the interests of a person who no longer exists. Philosophers supporting the surviving interests thesis have sought to elaborate on Dworkin"s reasoning or provide alternative justifications for why we ought to consider the patient as an extension of the prior person. Andrea Ott (2009) reasons that even if we identify with our psychological existence, we are nonetheless constituted by our bodies, that is, our human organism. Choices made by "Margo, the person," are also, derivatively, the choices of "Margo, the human." This leads her to claim that what "Dworkin is getting at is the intuition that the person and her organism are connected in a close enough way to generate such concern and respect for the life as a whole" (2009, 41).

There is something compelling about this, even if one doesn't accept Ott's account of personal identity. Even if we identify primarily with our mind in thought experiments about extraordinary scenarios such as mind swaps, our body remains important to our self-conception (DeGrazia 2005; Schechtman 1996). There is something intuitively problematic about treating "Margo, the person," and "Margo, the severely mentally impaired human" as moral strangers, as though Margo had always been severely

mentally impaired rather than spending most of her life as a person with complex goals and values.

David DeGrazia (2005, 79–84) and Mark Kuczewski (1999; 1994) describe surviving interests in terms of personal narratives. They adopt closely related versions of Marya Schechtman's (1996) account of personal identity, wherein our identity (not numerical identity, but identity in the sense of character) consists of our self-told life stories. Even though severe mental impairment may deprive Margo of her ability to continue developing her self-told narrative, the narrative that characterized "Margo, the person," involves Margo's life post-personhood. DeGrazia's point is that there is a strongly subjective component to whether our lives have gone well and that this is informed by the goals and values that matter most to us. Depending on Margo's goals, the factual events of her life might be construed as the life of a person who successfully secures what is most important to her or as a life characterized by efforts that are wasted in failed pursuits. Even though the capacity to tell this narrative requires personhood, the narrative is affected by the whole of Margo's life, including her post-personhood existence. As Kuczewski notes,

> "*People are their stories, not just their psychological states or their biological sensations. One's death forms an important part of the story. . . . The surviving interests of a person are interests in seeing the story carried out in a way that is meaningfully related to how it has proceeded up to the loss of decision-making capacity*." (Kuczewski 1999, 33)

For DeGrazia this narrative is constructed by the person herself (2005, 84–87), but for Kuczewski (1997, 135–38) it is a communal project between the person, her

community and their personal and institutional memories. Rather than taking moral authority as an act of independent self-creation, on Kuczewski's communitarian model, the advance directive serves to preserve the person's decision in the community's institutional memory: "The body that belongs to the incompetent patient at t_2 is in some sense 'mine' because other persons will call it by my name and make what happens to it a paper of the story they tell about 'me'" (1997, 13). As such, Kuczewski's is concerned for the advance directive as a means to the more important end of continuing the person's self-conceived narrative: "The continuing of the person's self-conception of their narrative self-construction is what makes the person the same person, not some act of will that is encoded in an instructional directive" (1999, 35).

Similarly, Dworkin is concerned for autonomy rights as a means of securing the opportunity to develop and meaningfully pursue "critical" interests—the nonexperiential but deeply held, goals and values that give our lives purpose (1993, 196–206). As with Kuczewski's communal concept of narrative, this suggests the possibility of a divergence between the patient's personal integrity and the precise terms of her advance directive. Nonetheless, an advance directive has relevance beyond being evidence of the person's state of mind at a fixed point in time. Advance directives are made within a legal and social context wherein they may function as a means of deliberate self-authorship, and the mere existence of the directive could serve to define one's critical interests or shape one's narrative in the communal mindset.

It is worth noting that the progressive nature of dementia itself complicates the stance taken by many proponents of the surviving interests thesis. Some who advocate a surviving interests stance regarding the advanced stages of dementia may nonetheless

alter their position once the patient's mental deterioration is so profound that even the marginal continuity created by minor familiarities, familial relationships or shared fragments of memories is completely extinguished. Allen Buchanan and Brock (1989, 154–57; see also Buchanan 1988, 282–92) limit their account of surviving interests in such a manner (requiring a minimal psychological continuity for personal persistence), and Kuczewski's communal understanding of narrative suggests that surviving interests may become marginal if the patient's former social relations are extinguished completely. By contrast, the emphasis that Dworkin (1993, 230) and Ott (2009, 41) place on the need to consider the patient's life "as a whole" appears to warn against limiting her critical interests, even under such extreme loss of psychological continuity.

4. The Experiential Claim

4.1 Rearticulating the Other Self Thesis

The accounts that warrant the "surviving interests" label posit a variety of conceptions of personal identity and of the relationship between "Margo, the person," and the human that survives her. Rather than disassembling each of these accounts and analyzing them piecemeal, I accept them because of the artificiality of judging "Margo, the post-personhood human," as though she were morally unrelated to "Margo, the person." Philosophical accounts of personal identity are drawn largely by observing the circumstances in which we feel prudential concern. Many people *are* prudentially concerned for their post-personhood existence, even when they are aware of the psychological change that will occur, and the surviving interests accounts explain why that concern may be rational. Whether we view "Margo, the post-

personhood human," and "Margo, the person," as two individuals or one, they are closely morally related, such that they have a legitimate moral interest *in each other*. This can work in either direction. Not only does personal integrity remain relevant to "Margo, the post-personhood human"; it is also part of the prudential interests of "Margo, the person," to ensure that her future post-personhood self is well provided for.

So where does that leave us? "Margo, the post-personhood human" is not a moral stranger to "Margo, the person," and so whatever the intuitive repugnance of killing her, it can't arise from that. Nonetheless, the idea that "Margo, the post-personhood self," might be a moral stranger to "Margo, the person," is difficult to dismiss as groundless prejudice, and the intuition isn't limited to those who downplay the relation between the preimpairment and impaired selves (e.g., Dworkin 1993, 228–29). Similarly, it doesn't seem quite adequate to rearticulate our intuitive concern as arising from some moral rule that is external to Margo, such as a general moral principle against passive euthanasia of a contented human—our concern is *for* Margo, not merely triggered by her. Insofar as this concern has a rational basis, it lies in the implication that Margo's current interests should be wholly subjugated by the pursuit of her former goals and values. On this interpretation, the concern that motivates the other self thesis is that Margo's life has moral worth of a kind that is not extinguished by the antipathy that she previously held toward such an existence. That is, Margo, in her post-personhood state, attributes value to her life in such a manner that she has serious moral worth that is independent of "Margo, the person."

A hypothetical case example may assist in more precisely defining my concern. What if a patient, "Bill," had for his entire adult life kept an unwavering belief in the moral

and spiritual value of flagellation, wherein he would arrange to be savagely whipped as a means of suppressing his material desires and building his character? In order to ensure that his valued practice is not disrupted by the onset of dementia, Bill makes an advance directive demanding that his future mentally impaired self be savagely whipped each week, and he finds someone willing to carry the whipping out. The example is absurd, but that is the point: its purpose is to show that the absurdity of the request *matters*. For simplicity, let's assume that there is no self-deception going on, and that Bill deeply values this practice far more than the experiential suffering it causes. The advance directive isn't demanding anything that goes beyond what we would allow Bill to do if he was choosing contemporaneously and in a mentally competent state of mind. That is not to suggest that the surviving interests thesis would treat Bill's well-being as a triviality. But, as Dworkin (1993, 201–6) argues well, a life in which one successfully pursues one's central nonexperiential goals is a better life than one in which one sacrifices those goals for experiential contentment. Say that Bill's commitment to flagellation is as central as Margo's hostility to extended advanced dementia and presumably of comparable importance to his personal integrity. It seems beyond absurd that we would view the advance directive as legitimately authorizing what is, essentially, torture carried out on a vulnerable and mentally impaired patient. We could create similar, admittedly far-fetched, examples by replacing the advance directive with one that contains instructions that are degrading or exploitative; say where the person requests that her future impaired self be spat on or (to fulfill her preimpairment commitment to protesting the global inequalities in availability of healthcare) be denied all pain relief. The point is not that these are realistic risks that require protecting against but that the pursuit of personal integrity through our future, severely mentally impaired, selves is subject to limits that do not apply when making the same decisions contemporaneously. Of course, these all

involve the infliction of positive abuse. That in itself might be a significant moral distinction from the withholding of life-saving treatment under heavy pain relief. But that simply emphasizes that the patient's post-personhood interests *matter* and that they matter in a way that is not predetermined by the relevance placed on them by the person she once was. The impaired patient has moral worth that is *independent* of her preimpairment goals and values and that limits the authority of her preimpairment instructions.

I will now try to rearticulate the concern that underlies the other self thesis. Whatever the value of the patient's commitments prior to her loss of personhood, it is not the only value that matters. The patient has moral worth independently of the person she was. The concern at the heart of the other self thesis is that by sacrificing a patient's experiential interests to implement her advance directive, we are not respecting her independent moral worth but rather are treating her as a mere vessel for the interests of the person she once was. If we are to respect the patient's interests both in her own current moral worth and in the person she once was, we need to look beyond the face of her advance directive. We must inquire whether the importance of the advance directive's contribution to the patient's personal integrity outweighs that of the experiential interests that would be sacrificed through withholding treatment.

In making this comment, I feel obliged to address a passage from the proclamation that has become known as "The Philosophers' Brief":

> "*A person's interest in following his own convictions at the end of life is so central a part of the more general right to make "intimate and personal choices" for himself that a failure to protect that particular interest would*

> *undermine the general right altogether. Death is, for each of us, among the most significant events of life*." (Dworkin, Nagel, Nozick, et al. 2007, 492)

I agree strongly with this declaration. My concern is not that too much importance is being placed on a person's interest in "in following his own convictions of the end of life" but rather that we are misconstruing that interest. In the context of the other self thesis, withholding life-extending treatment would not promote the patient's *own* convictions but rather her *previous* convictions. By sacrificing the things that the patient *now* cares about for those that she *once* cared about, we risk foisting onto her an inauthenticity that runs contrary to the interest that Dworkin and his coauthors urge us to protect.

If we are to accept that our interest in personal integrity continues past our loss of personhood, we must also accept the implication that we are not *just* persons. Underlying the importance that "The Philosophers' Brief" gives to convictions is an interest in living in accordance with one's own attribution of value. For those with the capacity for personhood, their "attribution of value" is reflected in their deeply held goals and values. They attribute value not only to particular experiences and objects but to ways of life. The claim to personal integrity requires that we respect this attribution of value, allowing those with the capacity for personhood to shape their lives in a manner consistent with their deeply held goals and values (Feinberg 1984, 52–97; VanDeVeer 1986, 124–27; Kleinig 1983, 27–30; Brock 1988b; Scoccia 1990). Those who lack the capacity for personhood are incapable of evaluating complex goals and values and of "shaping" their lives in the manner required for personal integrity. However, they can attribute value to their experiential interests, such as happiness or contentment, independently of their preimpairment views. This attribution of value

cannot, strictly speaking, be the basis for an interest in "*personal* integrity."

Nonetheless, the ability to imbue one's life with value is not an "all-or-nothing" capacity, and nor is it the sole domain of persons, or even of humans (Singer 1993; Taylor 1989, 27–-30). Similarly, nonpersons are entirely capable of *withholding* value from complex goals and values. It would be absurd to suggest that simply because a duck is not mentally capable of evaluating the great English hunting tradition, we can't be confident that the duck values its experiential interests more than it values that tradition. From the duck's incapacity to evaluate the tradition we may safely infer that the duck does not value it—we have no need to look for some earlier miraculous display of avian mental competence in order to ascertain the duck's attribution of value. It is at least as absurd to refuse to credit mentally impaired humans with substantial capacity for attributing and withholding value.

Respect for personhood is just one aspect of the broader principle of respect for the independent moral worth of other sentient beings. Insofar as a mentally impaired patient is capable of imbuing her life with value, we are obliged to respect that value. While this is not a claim to *personal* integrity per se, it is respect for integrity nonetheless—a demand that we determine the patient's good by reference to her own attribution of value. As such, respect for the patient's "life as a whole" requires that we not only consider the goals and values that informed her advance directive but that we also observe how her attribution of value has changed since that time and respect her interest in living and dying in a manner consistent with that value.

4.2 Self and Surviving Interests

In seeking to reconcile the "surviving interests" and "other self" accounts, I have made the following two claims: that the post-personhood patient has serious moral worth that is independent of the person that she once was and that the post-personhood patient is part of the life of the person she once was, such that her way of living, as well as the manner of her death, influences the integrity and value of that person's life as a whole.

These two claims are not conceptually contradictory; to the contrary, taken together they mirror some common attributions of moral worth. For example, I have moral worth that is independent of my status as a member of a family, while at the same time, I *am* a member of a family, such that my way of living is relevant to the well-being and value of that family. My claim to independent moral worth does not preclude me from having obligations to my family, and the severely mentally impaired patient's independent moral worth does not preclude the person she once was from having good moral claims over her way of life.

Nonetheless, the notion of holding moral worth independently *of one's own former self* requires some explanation. In accepting the general thrust of the surviving interests account, I have sought to avoid the Parfitian claim that we are numerically different individuals from our future selves. Rather than describing a moral conflict between different individuals, the claim that the severely mentally impaired patient has independent moral worth should be understood as a claim about "the good life" for an individual, as matter of viewing her life as a whole. In particular, it is a claim about the extent to which personal integrity determines "the good life" for an individual.

I outlined the appeal to personal integrity in the second section of this paper. Ordinarily, so the appeal goes, the good life for us is one lived consistently with our

own deeply held goals, values, and convictions, that is, it is a life that is consistent with our own attribution of value. Following this line of reasoning, a patient's informed choices are the ultimate indicator of "the good life" for that patient: the patient's own attribution of value is the authoritative arbiter of her life's independent moral worth. In any conflict between autonomy and well-being, autonomy must always win out, as the patient's well-being only matters to the extent that she values it (insofar as we are talking only of the patient's self-oriented concerns). Under *my* account, the patient's preimpairment attribution of value is not her life's *only* source of value, independently of other persons. In saying that a post-personhood patient has moral worth independently of the person she once was, we assert that her current (post-personhood) experiences can, if positive, imbue her life with value *regardless* of her preimpairment attribution of value. That is, in settling a conflict between the patient's preimpairment autonomy and her well-being, we ought to take into account two competing attributions of value: the patient's preimpairment goals and values and the patient's post-personhood appreciation of positive and negative experiences. Where the patient's directive is only of moderate importance to her preimpairment attribution of value but would be catastrophic with regard to her post-personhood attribution of value, we have good reason to subjugate the patient's autonomy in protection of her well-being.

The "surviving interests" account cannot by itself adequately explain why moral limits should apply to the fulfillment of goals that deeply characterize a patient's preimpairment personhood where that fulfillment would be permissible were the patient still competent—that is, it cannot explain why flagellation or painful but safe electric shocks could be delivered to a competent and consenting individual but not to a severely mentally impaired human under their instructions by advance directive.

Extrinsic explanations, such as potential harm to the reputation of the medical profession, misconstrue the target of our moral horror by making it seem as though the patient is only incidental to the grounds for her own protection. My account explains these limitations as a feature of the patient's own interests and thus provides an initial basis for doubting the blanket moral authority of advance directives—not by denying the appeal to personal integrity but by constraining it with a consideration of the patient's ongoing attribution of value. The patient is not *just* an extension of the person she once was but imbues her life with value and moral worth through her experiences of happiness and contentment.

5. Resolving the Conflicting Claims

I have argued that the moral authority of an advance directive ought to be constrained by an obligation to respect the independent moral worth of the patient in her severely mentally impaired state. This obligation becomes the fundamental question to be asked when determining whether an advance directive refusing treatment should be implemented: is the advance directive compatible with respect for that patient's worth or does withholding treatment here instead implicitly reduce the patient to a mere extension of the person she once was?

Some possible instructions could be so horrific that they could never be carried out while respecting the patient as the bearer of serious moral worth—for example, it would be out of the question to impose significant and unnecessary suffering or deliberate degradation. To carry out such instructions with respect to a severely mentally impaired patient would entail such abject cruelty that they could only appear warranted by viewing the patient entirely as an extension of the person that the patient

once was rather than as a being with serious moral value in her own right. The patient is capable of experiencing suffering but incapable of understanding the reasons for which we impose it, leaving her tortured and confused. The patient's ignorance and resulting fear magnify the harm imposed by such suffering beyond any plausible benefit. The withholding of life-extending treatment, if done with negligible suffering, is not inherently such a case.[7] The compatibility of passive euthanasia with respect for the patient's worth depends on our resolving the conflicting claims of the values protected by the advance directive (as per the surviving interests account) and the patient's ongoing experiential interests.

In resolving these claims, we can begin by noting the obvious: that many patients aren't as happy as Margo and aren't capable of such vivid experiential interests and that not all advance directives are equally vital to personal integrity. While such cases are less philosophically interesting, it is worth noting that there are "easy" cases as well as moral dilemmas such as Margo's. For example, there are cases in which the patient's unhappiness leaves us with no reason to override her advance directive and in which a patient's advance directive is grounded entirely in an ultimately misplaced fear that her post-personhood life will characterized by suffering. Elsewhere, a patient may be the subject of concern simultaneously under both the "surviving interest" and "other self" accounts. We face the dilemma of choosing between two moral concerns that would otherwise be morally essential: a person's interest in imbuing her life as a whole with meaning and her moral worth as a sentient being in pursuing a life of positive experiential interests. There is no need, as we have seen, to construe this as a

[7] The boundaries of what constitutes "negligible suffering" may prove to be a critical moral concern.

contest between two separate moral beings, nor is the dilemma adequately characterized as a trad- off between personal integrity ands well-being. Rather, we are concerned with how to interpret an individual's interest in living a life that embodies her attribution of value, under circumstances where her present and ongoing attribution of value consistently differs from that which characterized her as a person. Dworkin directly addresses this conflict: "There is a conflict between Margo's precedent autonomy and her contemporary experiential interests if she is still enjoying her life, but there is no conflict with her critical interests as she herself conceived them when she was competent to do so" (1993, 230).

That is, there are two attributions of value by Margo that warrant our moral concern, but only that attribution associated with "Margo, the person," is consistent with Margo's efforts to give her life, as a whole, meaning through the pursuit of her critical interests. Nonetheless, leading a meaningful life is not an all-or-nothing matter. Most people do *not* get the opportunity even to make an informed and relevant decision about the final stages of their life, let alone have those plans implemented. They die through accident or unexpected illness, or before they have given thought to their mortality, or under circumstances in which they feel deeply alienated by all possible choices. Many people consider deaths of this kind to be particularly tragic, despite their frequency. A large part of this extra tragedy comes from the loss of the opportunity to follow one's goals and values in death. Yet we often have little difficulty in reconciling this tragedy with the view that the person's life had great meaning when taken as a whole.

What the surviving interests thesis makes clear is that there is a certain type of person for whom control over the final stages of her life is utterly crucial to the meaning and

value she finds both in her death and in her life as a whole. The kind of cases I have in mind here are those where a person's life is characterized largely by self-sacrifice and self-deprivation that only become valuable to that person through their contribution to some greater goal or project that spans to the end of the person's life. Religious and ideological projects provide some of the clearest examples. The Jehovah's Witness's refusal of blood transfusions is nonsensical, even to the person herself, without the context of that person's larger project of living in accordance with a particular religious code. More importantly, the same can be said for the rest of the efforts and sacrifices that go toward "following her religious code." By forcing a blood transfusion upon the Jehovah's Witness, we deprive all of those efforts and sacrifices of that necessary context. Insofar as the person's life is characterized by the pursuit of the life project of "following her religious code," we trivialize her life by overriding her refusal of treatment.

This quality is not unique to moral and religious commitments. Rather, it arises because such prescriptions go beyond directing individual choices to set out a conception of the kind of person one wants to be and the kind of life one wants to live. These are commitments concerning the whole of one's life, drawing one's past and future choices together into a unified project. We could replace the example of the Jehovah's Witness with that of a quintessentially independent person who views her self-sufficiency as part of a broader goal of living an independent life in which she is not to be reliant on others financially or for day-to-day personal care. By the "quintessentially independent person," I mean someone for whom independence is a project that is embodied in her broader way of life: self-sufficient independence is a project that she has strived for and that shapes her choices and her self-conception. That is, she attributes value to the overall project of independence itself, as something

distinct from the choices that comprise it and the satisfaction she experiences through being self-sufficient. As with the Jehovah's Witness's religious practice, this means that the meaning and worth of her prior efforts are contingent on the success of the overall project. The quality of the person's final years can reshape the worth of her life, as measured against her own attribution of value.

These are extraordinary projects, representing one extreme with regard to differing views of what makes a life meaningful. The "quintessentially independent person" is not just someone who deeply values independence—in fact, it is not strictly necessary that such a person enjoy the minutia of such independence at all (though it would be a truly bizarre goal were that the case). The defining feature of such life projects is that the person attributes value to the "greater goal" of the life project as a whole rather than to the goals and values that comprise that greater goal. Certainly, the concept of a life where one commits oneself to the accomplishment of life projects, making sacrifices if necessary, forms one common concept of the good, and one that is strongly conducive to enabling the overriding emphasis that the surviving interest places on the fulfillment of a person's prior goals and values. But we may contrast it with another concept of the good life, one in which we find nonexperiential value in things enjoyed for their own sake. This has become a cliché with regard to the arts, as in the concept of the musician who is dedicated to music "for its own sake" or the playwright who writes plays "for the sake of good theater." To the person who has such a conception of what constitutes a meaningful life, the idea of attributing value not to her immediate commitment to art but to the greater goal of "living a life characterized by artistic achievement" may seem hollow, perhaps even self-defeating. A person who holds this concept of a meaningful life may ask what possible benefit there might be to being "a quintessentially independent person," above the value one

finds in one's immediate commitment to independence. Similarly, we can distinguish between the life project of the Jehovah's Witness who follows a religious code and that of the person who finds value in the minutia of religious commitment, such that her religious practices have meaning to her regardless of whether they are steps toward the accomplishment of a greater goal.

The latter type of "meaningful life" is not so malleable in its final stages as would be presumed by an overriding emphasis on the person's surviving interests. This is because the person has already taken value from her commitments in a way that cannot be undermined through later incapacity. If there is value in my immediate commitment to composing music, and I spend a time doing that, I do not lose that value if I cease composing music. I would, all else being equal, have a *better* life were I able to compose music for fifty years rather than compose music for thirty years and then lose the capacity to do so owing to advancing dementia, but the same would apply were my loss of capacity caused by death, pursuant to an advance directive, rather than dementia. Nonetheless, it is conceivable that I might take composing music to be so central to what I find good about life that I cannot imagine myself living a worthwhile life without it (from Kadish 1992, 871–888). My motivation, in this case, is not the completion of a life project but the prevention of a way of life that I do not value. Implicit in this is the denial that my future satisfaction has moral legitimacy—that is, it implies that such a life lacks serious moral worth because I do not presently value it, regardless of my attribution of value at that future time.

Where this type of motivation comprises the sole reason for the advance directive and at the same time euthanasia will deprive the patient of a happy life, I cannot reconcile euthanasia with respect for the patient's independent moral worth. My concern is that

euthanasia in this context seems to involve a kind of contempt for the moral worth of the impaired patient; that rather than being grounded in the pursuit of some vision of the good life, it is primarily concerned with the *denial* that value could be found in a severely mentally impaired life. Here, Euthanasia seeks to promote the patient's overall interests by *preventing* a way of life rather than promoting one. Euthanasia cannot restore or extend the patient's ability to lead a life that is independent of others or to write fine literature or do any of the myriad things that are unavailable to the severely mentally impaired. Instead, it is grounded in the notion that the patient's overall life is richer if we prevent her from living those ways of life that are open to the severely mentally impaired. That is, being unable to live the way of life that was central to her prior to advanced dementia, the patient cannot go on to discover a different way of life that has serious moral value. Where the patient *has* gone on to imbue her postimpairment life with great value through her ongoing happiness, this assertion—that the good life for the patient can only be pursued through her preimpaired goals and values—is akin to a statement of contempt for the mentally impaired patient's attribution of value.

These are moral extremes rather than being representative of actual people making advance directives. The very act of making an advance directive may amount to a declaration of a life project directly concerning the final stages of one's life. But it may also be motivated by fear an inability to imagine oneself with different views, prejudice against the mentally impaired, concern for one's close relationships, or the myriad of emotions and motivations that our mortality provokes. Most people, as Dworkin states, "want their deaths, if possible, to express and in that way vividly to confirm the values they believe most important to their lives" (1993, 211). But the amount that they invest in the final shape of their life, as well as what they stand to

lose if we override their advance directive, varies between individuals. This human diversity demands that we look beyond the face of the advance directive to determine the role it plays in that individual's life.

I cannot hope to provide a comprehensive stipulation of the factors that would render an advance directive refusing treatment morally just or unjust. Moreover, such a stipulation may not be advantageous. Even when providing a loose distinction of the kind I have given, there is a risk of encouraging others to see the relevant principles in terms of a set of rules governing when medical staff should be bound by an advance directive. This is misleading, as it makes out the constraint to be a far more legalistic matter than it is. The moral authority of the advance directive is constrained by nothing more than an obligation to respect the moral worth of the patient to whom it is applied. My concern so far has been to argue that such advance directives can, on occasion, lead to moral hazard and that we ought to look beyond the face of the advance directive. The actual means taken to safeguard the patient's interests do not need to take the form of a universal and comprehensive policy stipulating the circumstances in which advance directives should be implemented. Moreover, I fear that such a universal policy would risk an unwarranted extension of paternalism, leading to the patient's own attribution of value being replaced with that which underlies the policy. In the next section I suggest that our obligation to respect the patient's independent moral worth may be most appropriately safeguarded though the participation and oversight of a substitute decision maker appointed by the patient herself. I make this suggestion without consideration of the practical matters of politics and law, such as the effect on overlitigation and the feasibility of enforcement. For this reason, and given the limited space available, my comments in the next section should not be taken as anything more than a preliminary suggestion. Nonetheless, having argued for a

constrained use of advance directives to request passive euthanasia, I feel obliged to give some indication of how this constraint might be relevant to the practical implementation of such advance directives.

6. Implications and Shortcomings

On this articulation, the other self thesis does not imply that we should never implement an advance directive refusing life-extending treatment in the event of severe mental impairment. In that manner it is narrower than at least the versions put by Dresser (1994; 1995) and Robertson (1991). Nonetheless, it implies that we cannot know whether an advance directive is just without considering both the reasons the person had in making the advance directive and the extent of her current enjoyment of life. This is problematic for the practice of advance directives in the United States and as envisaged by Brock (1988a) in his description of advance directives as "performative utterances." Under that practice, the means by which advance directives carry out their function of protecting personal integrity is that once properly executed, the advance directive is legally binding (Bogdanoski 2009; Cantor 1993; Samanta and Samanta 2006). By describing advance directives as "performative utterances," Brock is noting that they are not simply *requests* regarding future treatment. They are rather instructions that are to be evaluated with regard to whether they fall within the social and statutory authority invested in that practice and not in relation to the adequacy of the person's reasons in that particular case. If the concerns I have raised in this paper are warranted, this existing version of the advance directive is poorly suited as a means of determining whether a patient's death should be hastened. A process whereby we implement the letter of a properly executed advance directive is well suited to a scenario where one's moral authority is unfettered by a need to consider values other

than one's immediate goals and values. Our moral authority over our future, post-personhood, treatment decisions is constrained by the independent moral worth of our future post-personhood life. This suggests a need for a kind of supervisory discretion that is at odds with treating the advance direction of passive euthanasia as a performative utterance.

Nonetheless, such decisions are not well suited to judicial oversight either. The moral authority of an advance directive refusing life-extending treatment turns on a great deal of information that is extraneous to the person's instructions. In particular, it involves the goals and projects that motivated the person to create the advance directive and the role that these goals and projects play in the person's life, as well as how overriding the advance directive will affect the meaning and value of the person's prior efforts. A judicial body would face insurmountable evidentiary difficulties if it were charged with assessing the justifiability of an advance directive refusing life-extending treatment in terms of the considerations that I and others have raised, if such an advance directive was to retain the same form as that of advance directives in general.

Having said that, I believe that decisions regarding euthanasia require us to reenvisage what an advance directive should be and how it should operate. An advance directive refusing life-extending treatment should combine the patient's instructions with what is effectively an application to have those instructions carried out. The advance directive should serve as a source of information that others can use to situate those instructions in the context of the person's life and projects. The advance directive should make out the person's case for withholding treatment, explaining how her commitments or convictions would be undermined by an extended life of severe mental impairment. For example, where refusal of treatment is vital to a patient's

moral code, the advance directive should record not only the patient's moral objections but should also detail the manner in which this moral code has permeated the patient's overall life, the self-sacrifice the patient has made in pursuing those ideals, and the loss that the patient will endure if treatment is imposed. This notion of an extensive advance directive, giving context for the instructions contained within, is certainly not new to bioethics (see, e.g., Emanuel, Danis, Pearlman, et al. 1995), though it has greater urgency in a context where another party is tasked with looking beyond the face of the patient's bare instructions. Such advance directives are typically envisaged as being produced through a process of consultation with the patient's doctor and, if necessary, others who through training will know what considerations the patient ought to take into account.

Even with such information being supplied within the body of the advance directive itself, the presence of a supervisory body introduces another layer of paternalistic intervention. Where this body is called on to assess the moral worth of the projects protected by the advance directive and of the patient's ongoing experiential happiness, there is a potential for the supervisory body to unjustly substitute their own values for those of the person for whom it decides. This risk is present with any supervisory body, but some are more prone than others. While a judicial body may be most reliably bona fide in its attempt to avoid such substitution (if that is demanded of it in the law), it is poorly placed to adopt and apply the patient's own values. This by itself may provide good reason to favor a system whereby the maker of the advance directive appoints a limited guardian with power of attorney regarding the advance directive over a system of direct judicial supervision. Such guardians have been discussed as alternatives to advance directives, on some occasions for similar reasons to those I have cited as constraints on euthanasia (Gedge 2004, Lynn, Teno, Dresser, et al. 1999).

Elizabeth Gedge (2004) argues that advance directives fail to capture the *process* of agency and self-governance; we make decisions and form our identities in large part through interaction and dialogue with those close to us, and by appointing such people as our substituted decision makers we may better approximate the choices we would have made. I would add to this the observation that agency and autonomy are not ordinarily pursued through choices that are fixed in time; we reconsider our choices as our goals and values change. To adopt the language of narrative theorists, advance directives reflect our story as it was at that particular point in time, whereas our narrative is more accurately something that progresses and adapts to changing capacities and circumstances. A substitute decision maker may be better placed to assess one's choice in light of one's life story as ongoing, especially in light of changes to one's capacities and values.

An advance directive in this re-envisaged form serves as the primary tool available to assist the substitute decision maker by revealing the patient's goals and projects prior to advanced dementia and by providing a context for those goals and projects. My hope is that a substitute decision maker who is properly informed by the advance directive and who is sufficiently close to the patient to have concern for both her preimpairment goals and her current happiness may better approximate the process of agency while ensuring that the patient's moral worth is respected. Jeffrey Bluestein (1999) recommends a similar method of substitute decision making informed by an advance directive, though he sees its merits more as a means of securing the continuance of the person's narrative than as a means of adjudicating between competing interests: "Proxy decision-makers are to act as continuers of the life stories of those who have lost narrative capacity, and it provides a defense of the moral authority of advance directives that is immune to the loss of personal identity

objection" (1999, 20).

Blustein (1999, 30–31) recognizes the independent worth of the post-personhood patient through her experiential interests but appears to view this as a counterweight to the authority of the substitute decision maker. My hope is that the substitute decision maker may be charged with concern for the patient's interests as a whole, that is, both the person's surviving interests and the post-personhood patient's continuing happiness. Like Dworkin and Kuczewski, Blustein's account seems to envisage a particular kind of person—one whose life is characterized by life projects that are capable of being meaningfully continued through an advance directive. Even if we allow that some advance directives may represent the creation of exactly such a life project, this is too narrow a concept of substitute decision making to address the full range of motivations that may underlie a person's plans for her post-personhood future.

In this final section I have sought to give a brief practical context for the concerns I raise earlier in the paper. My main thesis, however, concerns ethics rather than legal theory. While philosophers have largely supported the use of advance directives to demand that treatment be withheld, I am advocating a more cautious approach. Our implementation of a patient's advance directive ought to be constrained by our respect for the patient in her current state. This respect obliges us to consider the value of the patient's life by reference to her current capacities and attribution of value. Thus, I have sought to rehabilitate the "other self thesis" by restating it as a claim about the patient's interests rather than as about her identity as possibly different and unrelated (or marginally related) individual now from the she was before. I have not sought to rebut the works that I have grouped together as the "surviving interests thesis," but

have opposed the absolutist conclusion often drawn from this account: that the patient's most important interests are always those that occupied a central role in the life narrative of the person she once was. Finally, I state once more that I agree with the major message of "The Philosophers' Brief"—namely, the urgency of the moral claim to have one's convictions followed at the end of one's life. The supporters of an unconstrained right to have one's advance instructions implemented do not have a monopoly over those principles. My concern is that the interests underlying that statement of principles should be extended to the patient's post-personhood life. As Dworkin (1993) puts it, we ought to respect the patient's life "as a whole." This requires that we acknowledge that the patient is not *just* the meandering extension of a person's life, but that she is capable of serious moral worth that is completely independent of that prior person.

References

Arneson, Richard. 1980. Mill versus Paternalism. *Ethics* 90 (4): 470-489.

Bluestein, Jeffrey. 1999 Choosing for others: The Problem of Personal Identity Revisited', *Journal of Law, Medicine and Ethics* 27: 30-31

Bogdanoski, Tony. 2009. Psychiatric advance directives: The new frontier in mental health law reform in Australia? *Journal of Law and Medicine* 16: 891-905.

Brock, Dan. 1988a. Commentary on 'The Time Frame of Preferences, Dispositions, and the Validity of Advance Directives for the Mentally Ill'. *Philosophy, Psychiatry and Psychology* 5 (3): 251-253.

---. 1988b. Paternalism and Autonomy. *Ethics* 98 (3): 550-565.

Buchanan, Allen. 1988. Advance Directives and the Personal Identity Problem. *Philosophy and Public Affairs* 17 (4): 277-302.

Buchanan, Allen, and Brock, Dan. 1989. *Deciding for others: the ethics of surrogate decision making*. Cambridge: Cambridge University Press.

Cantor, Norman. 1993. *Advance Directives and the pursuit of death with dignity*. Medical Ethics. Indianapolis: Indiana University Press.

Christman, John. 2004. Narrative unity as a condition of personhood. *Metaphilosophy* 35 (5): 695-713.

---. 2005. Autonomy, Self-Knowledge and Liberal Legitimacy. In *Autonomy and the Challenges to Liberalism*, ed. Christman, John and Anderson, Joel. Cambridge: Cambridge University Press.

DeGrazia, David. 2003. Identity, killing and the boundaries of our existence. *Philosophy and Public Affairs* 31 (4): 413-442.

---. 2005. *Human Identity and Bioethics*. New York: Cambridge University Press.

Dresser, Rebecca. 1994. Missing Persons: Legal Perceptions of Incompetent Patients.

Rutgers Law Review 46 (2): 609-719.

---. 1995. Dworkin on Dementia: Elegant Theory, Questionable Policy. *Hastings Centre Report* 25 (6): 32-38.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

---. 1989. The Concept of Autonomy. In *The Inner Citadel: Essays on Individual Autonomy*, ed. Christman, John. New York: Oxford University Press.

Dworkin, Ronald. 1993. *Life's Dominion*. New York: Knopf.

Dworkin, Ronald, et al..2007. The Philosopher's Brief. In *Bioethics: Introduction to History, Method and Practice*, ed. Jecker et. al. 2nd ed. Sudbury, Mass.: Jones & Bartlett Publishers.

Emanuel, Linda, et al 1995. Advance Care Planning as a Process: Structuring the Discussion in Practice. *The American Geriatrics Society* 43: 440-446.

Fagerlin, Angela and Schneider, Carl. 2004. Enough: the Failure of the Living Will. *The Hastings Centre Report* 34 (2): 30-42.

Feinberg, Joel.. 1984. *Harm to Self*. Vol. 3. 3 vols. The Moral Limits of the Criminal Law. New York: Oxford University Press.

Frankfurt, Harry. 1988. *The importance of what we care about: Philosophical Essays*. Cambridge: Cambridge University Press.

--- 1997. *Fragmentation and Consensus: Communitarian and Casuist Bioethics*. Washington DC: Georgetown University Press.

---. 1999. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Gedge, Elizabeth. 2004. Collective Moral Imagination: Making Decisions for Persons with Dementia. *Journal of Medicine and Philosophy* 29, 4: 435-450

Harvey, Martin. 2006. Advance Directives and the Severely Demented. *Journal of Medicine and Philosophy* 31: 47-64.

Kadish, Sanford. 1992. Letting Patients Die: Legal and Moral Reflections. *California Law Review* 80: 857-888

Kihlbom, Ulrik. 2008. Autonomy and negatively informed consent. *Journal of Medical Ethics* 34: 146-149.

Kirby, Michael. 1983. Informed consent: what does it mean? *Journal of Medical Ethics* 9: 69-75.

Kleinig, John. 1983. *Paternalism*. Totowa: Rowan and Allenheld.

Kuczewski, Mark.. 1994. Whose will is it, anyway? A discussion of advance directives, personal identity and consensus in medical ethics. *Bioethics* 8 (1): 27-48.

---. 1999. Commentary: Narrative Views of Personal Identity and Substituted Judgement in Surrogate Decision Making. *Journal of Law, Medicine and Ethics* 27 (1): 32-36.

Kuhse, Helga. 1999. Some Reflections on the Problem of Advance Directives, Personhood and Personal Identity. *Kennedy Institute of Ethics Journal* 9(4): 347-364.

Lynn, Joanne et al. 1999. Dementia and advance care planning: Perspectives from three countries on ethics and epidemiology. *Journal of Clinical Ethics* 10:(4) 271-285.

McMahan, Jeff. 2002. *The ethics of killing: problems at the margins of life*. New York: Oxford University Press.

Meyers, Diana. 2005. Decentralizing Autonomy: Five Faces of Selfhood. In *Autonomy and the Challenges to Liberalism*, ed. Christman, John and Anderson, Joel. Cambridge: Cambridge University Press.

Mill, John Stuart. 1977 (first in 1859). On Liberty. Ed. Robson, John. The Collected Works of John Stuart Mill. Toronto: University of Toronto Press.

http://oll.libertyfund.org/title/233.

Ott, Andrea. 2009. Personal Identity and the Moral Authority of Advance Directives. *The Pluralist* 4(3): 38-54.

Post, Stephen. 1995. Alzheimer Disease and the "Then" Self. *Kennedy Institute of Ethics Journal* 5 (4): 307-321.

Robertson, John. 1991. Second Thoughts on Living Wills. *Hastings Centre Report* 21 (6): 6-8.

Samanta, Ash, and Samanta, Jo.. 2006. Advance directives, best interests and clinical judgement: shifting sands at the end of life. *Clinical Medicine* 6: 274-278.

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornwell University Press.

Scoccia, Danny. 1990. Paternalism and Respect for Autonomy. *Ethics* 100 (2): 318-334.

Singer, Peter. 1993. *Practical Ethics*. Cambridge: Cambridge University Press.

Taylor Charles. 1989. *Sources of the Self: the Making of the Modern Identity*. Cambridge: Harvard University Press.

---. 2003. Autonomy, Duress, and Coercion. *Social Philosophy and Policy* 20 (2): 127-155.

Valerius, Jukka. 2006. On Taylor on Autonomy and Informed Consent. *The Journal of Value Inquiry* 40: 451-459.

VanDeVeer, Donald. 1986. *Paternalistic Intervention: The Moral Bounds on Benevolence*. New Jersey: Princeton University Press.

Young, Robert. 1989. Autonomy and the "Inner Self". In *The Inner Citadel: Essays on Individual Autonomy*, ed. Christman, John. New York: Oxford University Press.

Paper 7 – Problems with the easy justification

Abstract: In this concluding piece, I discuss the worst injustice of our current system of mental incompetence, whereby people 'qualify' for moral agency through a test of competence that is necessarily evaluative, but governed by values that are not open to adequate scrutiny. The consequence is that a small but important minority of people who suffer from mental illness are excluded from our 'society of persons', in which we attribute moral and rational agency to each other despite our ever-flawed mental capacities. Whilst the exclusion is formally in relation to only a small aspect of social interaction; most notably that of moral agency in relation to violent conduct, the resulting involuntary hospitalisation gives the effect of a much broader form of exclusion.

I argue in this paper that involuntary psychiatric treatment must *only* be carried out for paternalistic purposes. To do otherwise commits a congregation of injustices, from the division of doctors' duties wherein they make decisions on behalf of a patient but not *for* the patient, to the imposition of a discrimination similar in immorality to that of sexism and racism, and most importantly the social disenfranchisement of the mentally ill. Once again, and carrying on with the central thesis of this dissertation, the singling out of the mentally ill for non-paternalistic preventative detention is not the inescapably logical outcome of medical science, and we do our mentally ill members of society a grave disservice if we hide behind medical science when depriving them of their right. As I have argued throughout this dissertation, the scope of moral agency and the terms of inclusion in the society of persons is a social and moral decision – one informed by medical science, but not governed by it. In earlier papers I have argued that the pretence of value-neutrality

distracts attention away from the vital question of whose values it is that governs treatment and determines the criteria for personhood. By exposing the mental ill to non-paternalistic involuntary treatment we render them vulnerable to being stripped of personhood on the basis of values that are foreign to themselves and their subculture. I also argue that my account of patient autonomy makes considerable allowance for the paternalistic prevention of illness-induced violence, as a horrific usurping of a person's capacity for the construction and pursuit of a personal identity that is authentic to herself and her conception of the good. Nonetheless, I argue that the only moral outcome is to prohibit entirely any involuntary psychiatric treatment that is not governed by the interests of the person being treated: if we are to make decisions on behalf of a patient, we must make those decisions *for* that individual.

1. Preventative detention: the 'easy' justification for civil commitment.

1.1 An oversight in civil commitment

Involuntary psychiatric hospitalisation takes place as a part of the broader medical institution, in that its most basic and immediate function is the provision of medical care to those with serious illness. Nonetheless, the forceful imposition of psychiatric treatment runs so contrary to ordinary medical ethics, and in particular the informed consent doctrine, that it requires moral and legal justification beyond the benefit inherent in treating illness. Two such justifications regularly appear in the legal criteria for involuntary admission and in the written commentaries about the resulting system. The broadest of these is the paternalistic justification, whereby the patient's liberties are restricted in order to protect the patient's own interests. My main concern in this paper is the other justification: that of preventing the patient from *harming other people*. Both justifications may give rise to what can properly be described as 'preventative detention'. However, for ease of use when comparing involuntary psychiatric hospitalisation with relevantly similar systems of detention, I shall use the term 'preventative detention' to refer *solely to detention for the purpose of preventing a person from causing*[8] *harm to others*.

Laws on preventative detention vary considerably between nations, and my description of the legal and practical context concerns Australia, USA and Britain foremost (and by extension the many nations that share the British legal traditions on civil commitment)[9]. 'Preventative detention' here refers to a form of civil commitment,

---

8 Causation, in this sense, does not imply moral responsibility and includes such causal chains as when one person innocently 'causes' another to become infected with a contagious illness.

9 Typical examples (by no means exhaustive) of the mental health framework that characterise preventative detention in Australia, the UK and the USA are *Mental Health Act 1996* (WA) in Australia, *Mental Health Act 1983* (UK) in England and Wales, and *Florida Mental Health Act 1971* in the United States. Each requires psychiatric evidence of illness

requiring no criminal charges or formal record of violence. Preventative detention (as part of civil commitment) must be distinguished from the criminal defence of mental impairment, and the citing of mental illness in a plea of mitigation during criminal sentencing. In each such circumstance, a court may order treatment directly, or preferably (and with growing frequency) subject the person to the determination of a tribunal with more appropriate expertise to determine treatment needs (e.g Mental Health Act 1983 (UK) s2, 3, 37; Sentencing Act 1995 (WA) Part 8; Sentencing Act 1991 (Vic) Div. 3; Crimes (Mental Illness and Unfitness to be Tried) Act 1997 (Vic); Criminal Procedure (Insanity and Unfitness to Plead) 1991 (UK)). This practice has its own areas that need reforming, such as in jurisdictions where a criminal defence of insanity may subject the person to indefinite detention by detention (Sentencing Act (WA) 1991 s21, 24). However, for current purposes its most important quality is that it subjects the mentally ill to the encroachment of the state only in relation to those same behaviours and circumstances that attract police attention for other citizens. Diversionary orders for psychiatric treatment serve as non-punitive mitigatory responses to criminal charges, not as extraneous impositions that single out the mentaly ill for extra restriction. Most importantly, they render the mentally ill subject only to those forward-looking offences that govern the entire community – they are liable to prosecution (and hence psychiatric treatment) for reckless driving or firearms offences, but not to any vague charge of 'dangerousness'.

---

and a risk of harm that can not be adequately prevented through less intrusive means. Whilst the United States jurisdictions often have similar risk-based criteria to Australia and the UK, their more extensive body of case law has established that the harm must be much more imminent, and possibly more severe, than is required for involuntary hospitalisation in Australia and the UK (O'Connor v Donaldson, 422 US 563 (1975). United States law usually requires court hearings for non-emergency detention, whereas in Australia review of involuntary hospitalisation is conducted through relatively informal and brief tribunal hearings, usually without legal representation.

Preventative detention, however, serves as an alternative branch of the same commitment procedures that are utilised in imposing involuntary treatment for paternalistic reasons. In Australian jurisdictions, this bypasses the judicial system entirely (though review by quasi-judicial tribunal, as well as an appeal path culminating in judicial review, remains available) (Mental Health Act 1983 (SA); Mental Health Act 1996 (WA); Mental Health Act 1996 (Vic); Mental Health Act 2000 (Qld); Mental Health Act 2007 (NSW)). Crucially, detention here is *not* dependent upon lack of mental competence, but solely upon the co-existence of any mental illness and some risk of harm to other persons or property. That is, preventative detention is *risk*-based, not *competence*-based. There is no legal or logical inconsistency, therefore, in civilly committing a person because she poses a risk to others, and applying criminal (and moral) responsibility to that same person in the event that the predicted harm manifests. The moral premise for intervention is the minimisation of harm, not the absence of moral capacity.

It is fair to say that this second basis for involuntary hospitalisation has received a great deal less scrutiny than the paternalism function. During the past half century, psychiatric paternalism has been the subject of a sizeable volume of work by philosophers, legal commentators, sociologists and doctors (Papageorgiou et al. 2002; Dworkin 1986; Kleinig 1983, 1–31, 67–73; VanDeVeer 1986, 88–127; Ward & Savulescu 2006; Savulescu & Momeyer 1997; Failer 2002; Morse 1982). Preventing harm to people other than the patient, by contrast, seems to be viewed as the 'easy' justification; one that is simply too obvious to critically analyse. Exceptions to this are rare, and seem to come from those few radical authors that reject the entire concept of mental illness upon which such intervention is based (e.g. Szasz 2008; Szasz 2003).

During the great revival of broad philosophical liberalism during the 1980s, the landmark works on paternalism adopted a predominantly neo-Millian liberalism (Arneson 1980; Feinberg 1984, 53–71; VanDeVeer 1986, 88–127; Kleinig 1983; Reznek 1991). It may be that the Millian exception of the 'harm' principle, together with the fear of an expanding paternalistic state, has led the 'easy justification' for involuntary hospitalisation – i.e. prevention of harm to others – to avoid extensive scrutiny. Certainly, the lack of serious sceptical study of this branch of the civil commitment provisions is quite incredible compared to the enormous critical scrutiny of preventative detention when aimed at sex offenders and terrorism suspects (Carne 2007; Corrado 1996a; Gray 2004; Gray 2005; Kelly 2008; Lippke 2008; Robinson 2001; Kitai-Sangero 2009; Schoeman 1979). Preventative detention of the mentally ill shares almost all of the features that legal authors and ethicists find troubling in these cases - the bypassing of criminal processes, the merging of executive and judicial functions, reliance on evidence falling short of the 'beyond reasonable doubt' standard and the absence of a clear retributivist justification for the detention – while lacking either of the two features that might warrant singling this group out, i.e. serious criminal history or a significantly increased likelihood of gross violence (Allen 2005; Link & Stueve 1994; Robinson 2001; Dershowitz 1970; Slovic & Monahan 1995).

The burden of preventative detention is broader than the immediate detention itself. Only a small minority of people with mental illness are ever likely to be deemed dangerous in a manner that exposes them to detention, but the exposure to assessment for dangerousness, as well as the more abstract denial of legal equality, affects the mentally ill as a *group* (compare a regime of selectively detaining African Americans,

defended on the grounds that 'it only affects the dangerous ones'). Part of the burden of preventative detention is the absence of civil liberties, legal safeguards and individual rights that are ordinarily available to criminal defendants at the stages both of arrest and hearing, I do not address this point in detail, as many of these concerns have been dealt with extensively by legal academics in the context of preventative detention of sex offenders and terrorism suspects, but the gap in legal rights is considerable. Criminal defendants enjoy a series of rights that vary between jurisdiction but often include the right to cross-examine all witnesses who present evidence against them, legal representation at all stages of the process from police interrogation to judicial sentencing, the right not to have negative imputations drawn from refusal to participate in interviews (now severely limited in the UK), the exclusion of evidence that is more likely to be deceptive than probative (such as hearsay evidence), the criminal standard of proof (i.e. beyond reasonable doubt), and a check on the power of prosecuting authorities through a major bureaucratic division in the organisations responsible for policing and prosecution (e.g. Criminal Procedure Act 2004 (WA), Miranda v Arizona [1966], Criminal Procedure Rules 2011 (UK))**.** Most notably, it includes a prohibition on evidence of poor character (except to counter a claim to unusually good character) – the precise kind of evidence that informs preventative detention judgements (Evidence Act 1906 (WA) s8(1)(e)). Civil commitment procedures superficially allow a parallel set of safeguards, such as the right to appear before a tribunal and cross-examine witness evidence with legal representation, but in practice their utility is far more limited than that available to criminal defendants. In Australia, for example there is no legal review at the point of initial detention, allowing detention for weeks or months before tribunal review occurs, during which time the patient has no legislative basis upon which to cross-examine witnesses or examine the manner in which evidence has been obtained and

considered (e.g. Mental Health Act 1996 (WA); Mental Health Act 2003 (Qld); Mental Health Act 1993 (SA), Mental Health Act 2007 (NSW); Mental Health Act 1996 (Vic)). As a civil action the 'beyond reasonable doubt' evidentiary standard for criminal prosecutions does not apply. The sources and content of evidence informing psychiatric judgement are withheld from patients where their psychiatrist (usually the same person whose treatment order is being reviewed) states that knowing the information is contrary to the patient's health or relationships, or reveals information given in confidence, or personal information about another person, regardless of their importance to the psychiatrist's decision (Mental Health Act 1996 (WA) s161). Legal theorists examining preventative detention in other contexts have also noted its abandonment of key legal protections, including the need to demonstrate intent to commit the offending act, difficulty effectively cross-examining adverse evidence where courts rely upon abstract expert judgments, and the consistent and repeated demonstrations that psychiatric and other predictions of dangerousness are inaccurate, erring heavily towards over-prediction of dangerousness (Dershowitz 1970; Ashworth 2000, p.180; Gray 2004; Groethe 1977, 583–584).

In this paper, I argue that the programs of preventative detention, that we have for so long taken for granted as the least controversial basis for civil commitment, rest on morally perilous grounds. I do not necessarily reject the practice of preventative detention altogether: there may be a case for preventatively detaining some extraordinarily dangerous individuals. Instead, I doubt that a just society can reasonably place any extraordinary burden upon the mentally ill as part of such practices. That is, we cannot justify imposing upon the mentally ill a regime of preventative detention that we do not impose upon the general population.

Similarly, the criminal law has long included preventative components – licensing of drivers and guns, punishment of behaviours like drunken driving and acting drunk and disorderly – the policing of things that that do not necessarily cause serious harm to anyone but create a risk that such harm could occur. I have no argument against the imposition of treatment orders where a person is found to be mentally incompetent in breach of such laws, so long as she is not thereby deprived of the safeguards on state prosecution that are available to other individuals. In summary, I take no present concern with the imposition of some non-discriminatory policy of detaining sufficiently dangerous individuals, nor with the prohibition of dangerous behaviour, and I certainly do not oppose the imposition of treatment orders rather than mainstream imprisonment or punitive orders in the event that a person is found to be mentally incompetent in relation to criminally dangerous conduct or characteristics. I do, however, take issue with the singling out of the mentally ill for preventative detention. In effect, I believe that psychiatrists and medical institutions should hold a duty solely to the unwell patient, protecting them from self-harm and the unjust imposition of undeserved punitive sentences.

There is already extensive material analysing the extrinsic problems facing preventative detention (in the context of detaining past sex offenders), and in particular many authors have debated whether the risk of false positives should rule out such intrusive intervention (Carne 2007; Corrado 1996b; Corrado 1996a; Fairall & Lacey 2007; Gray 2004; Gray 2005; Groethe 1977; Kelly 2008; Kitai-Sangero 2009; Lippke 2008; Montague 1999; Morse 1996; Morse 1982; Slovic & Monahan 1995; Underwood 1979). Such criticisms are certainly relevant, but for now I put them aside. No matter how inaccurate our predictions usually are, and how great the damage done by false positives, there will be exceptional cases where we are completely certain that

a person is going to cause serious harm. Preventative detention is, by design, a response to exceptional circumstances, and so we should be interested not only in whether it is *typically* just but also whether it may be a just response where the risk of harm is extraordinarily certain and dire.

The abolition of preventative detention as a ground for civil commitment would be a substantial reform, but not necessarily so radical an alteration as it may initially appear. I suspect that paternalistic concern for those who may be driven to violent conduct may often provide an alternative basis for intervention, and one which provides a far more intuitive reason for sidestepping the safeguards enjoyed by criminal defendants – if the aim of such safeguards is to protect the accused, then it stands to reason that they ought not act as barriers to providing appropriate assistance to that individual. I do not pursue this issue at length, but merely raise it as an indication of the possible consequences of abolishing preventative detention. In particular, I illustrate how an argument for such a policy could *potentially* be constructed using the extended concepts of self and autonomy that I describe in the papers that comprise the first half of this dissertation. This would not be a mere swapping of labels for essentially the same behaviour. Limiting preventative detention to a strictly paternalistic policy, outside of which the mentally ill are entitled to all the civil liberties enjoyed by other citizens, changes the focus from risk to mental competence. As I have already noted, the two do not always overlap, making it possible for a person to be criminally prosecuted and civilly committed for the same set of events. Moreover, there is an injustice in exposing the mentally ill to different legal repercussions of an assessment of dangerousness than that applied to the general community, unless that deviation is supported by a difference in the person's capacity for moral responsibility and autonomy. Our dual system of civil commitment, whereby

the mentally ill are both treated and policed by the same institutions, corrupts the aims of our medical institutions by forcing them to act as instruments of social inequality. By replacing it with a system focussed solely on paternalistic protection of the mentally ill, where policing and risk-prevention apply to the mentally ill only in that manner which is acceptable when applied to the general community, we can achieve genuine ground-level reform of the way in which the mentally ill interact with medical institutions.

1.2 Why mental illness?

Most people who suffer from a mental illness at some stage of their lives are not dangerous, whereas most dangerous individuals fall outside the group targeted by non-paternalistic involuntary psychiatric treatment programmes (Lavin 1995). While the spectre of the murderous madman continues to loom within second-rate fiction, the reality of the minimal relationship between mental illness and violent crime has been known for long enough that it draws the question of why we continue to consider mental illness to be a characteristic worthy of preventative detention policy. Of course, out of the enormous range of mental illnesses recognised by psychiatry, most have little bearing upon civil commitment of any form. It may be more pertinent to point to studies that have identified a narrower range of mental illnesses and symptoms that magnify the risk of violent crime more significantly, e.g. 'threat and control override symptoms' that double the probability of violence, and stronger correlations with potentially psychotic illnesses, with schizophrenia in particular increasing the longitudinal likelihood of violent crimes by around four to six times (Fazel et al. 2009; Fazel et al. 2010; Link & Stueve 1994; Swanson et al. 1996; Swanson 1994). However, recent extensive studies of schizophrenia and bipolar have demonstrated a powerful increase in the proportional risk of violence where *substance abuse* is

simultaneously a factor, to the extent that the heightened risk of violence associated with mental illness may be primarily a consequence of greater substance abuse, rather than an intrinsic feature of the illness (Fazel et al. 2009; Fazel et al. 2010). However, the most important comment on the relationship between mental illness and violence may have been made by Groethe (1977), over 30 years ago. Violence (more specifically, violence of a kind that is institutionally policed) is rare behaviour, and even if a mental condition multiplies the risk of violence several times over, that falls far short of demonstrating that a person with that condition is *likely* to engage in violence.

In any event, there are conditions with far greater links to violence than mental illness, such as employment history, age and family situation (Robinson 2001). Not to mention the condition whose correlation with crime is so obvious and overwhelming that we view it as trivial: that of being a male, whereupon the FBI's national database of arrest statistics reveals that in 2010 74.5% of arrests for violent crimes in the USA were of males (Federal Bureau of Investigation 2010). Of course, even though preventative detention affects the civil rights of all persons within the class affected, it aims to only detain the small minority that buck the trend and are seriously dangerous *individuals*. But seriously dangerous individuals are certainly not unique to the mentally ill. If the justification for singling out the mentally ill for preventative detention is that of an exceptionally heightened risk of dangerousness, a *much* better case could be made for screening of unemployed young men in accordance with a program of preventative detention. I expect that most would find this intolerable, for reasons largely in accordance with the principles of the presumption of innocence, criminal intent and other legal doctrines restricting the power of the state in matters of criminal law**.**

There is a further danger in citing actuarial studies of mental illness and violence as a justification for preventative detention. Even as studies take account of the influence of gender and socio-economic variables, and more recently substance abuse, there remains a risk of 'double-counting' circumstances indicating dangerousness, where those circumstances are both directly observable and the result of mental illness. Say that a person holds strange beliefs that seem impervious to opposing evidence, and in accordance with those beliefs adopts a deeply angry disposition towards an individual or group, threatening violence against them. Such a person is likely to be equally dangerous regardless of whether these features are caused by mental illness or not – mental illness *explains* why the person holds these dangerous characteristics, but does not make them any more dangerous than if the same characteristics were due to some other cause. There is a risk of serious error then, in applying actuarial data to individual cases. Aside from our inability to know whether the individual is typical of the studied group, the use of actuarial data can falsely imply that the person is *more* dangerous than the sum of his characteristics would indicate, simply because the decision-maker is adding the actuarial risk arising from mental illness to the directly observed manifestation of that *same* risk. In fact, given the aforementioned rarity of violence amongst the mentally ill, mental illness is far better at *explaining* violence than it is at *predicting* it. The statistical near-irrelevance of mental illness itself provides something of an argument against preventative detention targeting the mentally ill: if the bald fact of severe mental illness tells so little, and the decision to impose preventative detention can only be justly made if it is based on a more direct examination of the person's behaviour, history and mentality, why shouldn't we detain anyone with those dangerous features, irrespective of whether they are explained by mental illness? Why should two people, who display near identical behaviour and views, and thus present the same risk of violence, be treated differently simply because

one person's characteristics are best explained by mental illness, and the other person's characteristics are best explained by local cultural norms?

However, our motivations for preventative detention have very little to do with any broad threat posed by the severely mentally ill. There are *some* occasions where mental illness leads a person to violence, and policies of preventative detention aim to prevent those few occurrences, even in the knowledge that the mentally ill are not substantially riskier than the general population. Given the haphazard nature of the political process, and the recent expansions of preventative detention, it may be more charitable to view the scope of preventative detention as being in a state of flux: so far, the political process has turned its attention to the mentally ill, terrorism suspects and sex offenders, but in the future it may turn its eye to other explanations for dangerousness. The more pertinent question, then, is whether the principle on which the mentally ill are targeted and assessed is a fair one.The problem with singling out the mentally ill is that it involves discrimination on the basis of what I describe as 'depersonalised traits', traits which describe the individual in objectified terms, making the consequent discrimination a form of disenfranchisement analogous to institutionalised racism and sexism.

2. Blameless detention and the tragedy principle

'Preventative detention' is, almost by definition, an *undeserved* response to some existing or past failing. The criminal law has preventative components such as reckless driving, or bringing a firearm into parliament, but the punishment for such offences is comprised of a mixture of goals, including the punishment of the person's failure to have appropriate regard for the safety of others, in addition to that aimed at deterring

both that individual and others from engaging in such dangerous behaviour. That latter component – deterrence - is punishment that targets *potential* conduct, i.e. something that might never eventuate. To the extent that detention is preventative, it is undeserved but imposed for the public good.

Preventative detention is then an *undeserved* burden upon those who are detained. Even where our prediction of future harm is accurate, in that the person *would* have caused harm but for the state's intervention, there is no relevant moral failing warranting detention at the time it is employed. This lack of a readily available retributive justification for preventative detention has greatly concerned legal theorists, though almost exclusively with regard to the detention of repeat sex offenders and suspected terrorists rather than the mentally ill, (Carne 2007; Corrado 1996a; Gray 2005; Gray 2004; Kelly 2008; Kitai-Sangero 2009). In Australia, the High Court rejected as unconstitutional a New South Wales statute that sought to establish a system of preventative detention targeting prior sex offenders (Kable v DPP [1996]). Whilst the Court's reasoning rested largely on a separation of powers doctrine, Justice Gaudron was explicit in his concern about the injustice of preventative detention (at 106-107):

> *'That is the antithesis of the judicial process one of the central purposes of which...is to protect the individual from arbitrary punishment and the arbitrary abrogation of rights by ensuring that punishment is not inflicted and rights are not interfered with other than in consequence of the fair and impartial application of the law to facts which have been properly ascertained.'* (Kable v DPP (1996) 189 CLR 51 at 106-107)

Perhaps for this reason, some advocates of preventative detention have sought to bring

it back within the scope of retributive justice by proposing that we view it as *pre-punishment*, justified by the moral failings we *expect* a person to make (Lippke 2008). I am immediately skeptical of the notion that this 'pre-punishment' can be grounded in moral desert – desert is inherently backwards-looking as it requires that we judge a person by her morally autonomous choice, rather than merely surveying the circumstances, traits and character flaws that promote the odds of her choosing to harm others. Moreover, it seems simply to be a convenient and not altogether truthful *excuse* for preventative detention, rather than a plausible statement of our reasons. If our concern was to ensure that wrongdoers are punished, we could do that with far greater accuracy by waiting until after the offence is committed. Our reason for intervening prior to the harm is quite clearly because we want to prevent the harm from occurring, and that is what one would expect to form the basis for our justification. However, even if we put aside all concerns about the punishment coming before the offending conduct rather than after, predicted harm doesn't give rise to the kind of desert needed to justify criminal punishment. When assessing the likelihood of future harm, the best we can do is conclude that an individual poses a *risk* or a *probability* of harming others. The problem is not the mere fallibility of prediction: any criminal justice system bears some risk that the innocent will be convicted, and over a sizeable number of cases it becomes almost inevitable that some innocent person will be imprisoned. The problem is simply that dangerousness does not imply guilt.

Say that we have a system of criminal justice with a consistent accuracy of 95%, such that out of a strictly representative sample of 100 convicted individuals we'd expect 5 to be innocent and 95 to be guilty. Now say that we have a remarkable means by which we can predict that people within an altogether separate group are 95% likely to cause

harm. These two groups are very different, and it is through overlooking this difference that Lippke (2008) errs in suggesting that the risk of false positives in preventative detention is no more troubling than a system of ordinary criminal punishment with a comparable error rate. Those subject to preventative detention cannot be separated into deserving and harmless individuals, such that we can decide upon an acceptable rate of false positives. If we were to take 100 people from the second group, i.e. wherein they pose a 95% risk of causing harm, we do not have 95 future offenders with 5 innocents hidden among them. *All* 100 are equally dangerous and equally innocent; we can't be treating 95 fairly and doing the other 5 an injustice because they all share the same basis for detention (membership of the relevant group). In fact, we have no guarantee that any of them will harm anyone – they all pose a risk of harm that is not actualised at the time of detention. To 'pre-punish' these people would be to ascribe to them a future that they don't yet have, as though it was already set in stone. None of them have a future in which they will harm others; only an unactualised possibility of such a future. The most we can say is that they are *dangerous*, that they are the *kind* of people who would harm others. A person does not deserve punishment merely because he or she is dangerous, and we cannot reconcile the gap between dangerousness and actual harm by comparing it to the risk of false conviction.

Retributive justifications for preventative detention, i.e. those which seek to punish a person for some future crime or moral failing, face daunting logical and practical barriers. But do we really *need* to identify some future crime or moral failing before we intervene? If we look beyond the recent expansion of preventative detention to its more traditional forms, i.e. quarantine and non-paternalistic involuntary hospitalisation

for psychiatric treatment, criminal punishment becomes an increasingly uninviting analogue. Non-paternalistic involuntary hospitalisation is most appealing where the patient is *incapable* of moral and criminal responsibility, or where moral responsibility is simply irrelevant to the reasons for intervention. For example, quarantine, in the context of containing the spread of a highly contagious and potentially deadly disease, doesn't raise the troubling issue of pre-punishment because it has no relation to any human wrongdoing.

Corrado (1996b) cites the relative lack of moral concern regarding quarantine as evidence that there is no moral requirement that detention only be imposed on retributive grounds. Corrado's comparison is a plausible one: whilst quarantine is not intended as criminal punishment, and no doubt involves more humane facilities and care, any physical detention involves a broad restriction of one's liberties and an almost complete denial of the ability to pursue the projects, commitments and relationships that ordinarily give one's life value. A similar comparison of quarantine to civil commitment was made by the United States Supreme Court in United States v Comstock [2010]. Yet there is no suggestion that those subject to quarantine in any way deserve that burden. If quarantine has any moral basis, it is in preventing great harm – the notion that some sufficiently dire potential harm (taking into account both the severity of the harm and the likelihood of its occurrence) warrants our taking extraordinary action at the cost of those unfortunate enough to present a risk of contagion.

Quarantine occurs in a context where harm through inaction is as morally troubling as harm through direct action. The state is not an innocent bystander: it has a positive

duty to protect the population from known and unwanted harms. In this sense, quarantine is an attempt to fairly divide the losses imposed by a tragedy. There is no 'deserving party' to whom the losses can be allocated; quarantine merely protects one group from suffering an undeserved loss, at the cost of imposing an equally undeserved (but presumably lesser) loss upon another group. This is detention without crime or punishment. There is no suggestion of moral failing and the aim is not to punish but to mitigate potential loss.

From the example of quarantine, we might suggest a 'tragedy principle': that where we have a positive duty to prevent or minimise harm, we may impose an undeserved burden upon one group of people if necessary to prevent some (sufficiently) greater and equally undeserved harm from afflicting another group. If so, we should expect this to apply as well to the prevention of serious violence, insofar as it is a significantly greater harm than that of involuntary hospitalisation or imprisonment. The tragedy principle implies that we should treat all involved as equals, in that we do not assume that the 'natural' distribution of burdens (i.e. without our intervention) has any particular moral authority. Those who would be free from any relevant burden without our intervention do not thereby have a right to remain unburdened when we distribute the losses involved. That doesn't mean that the state should always act to minimise the total harm; a distribution may be unfair because it places a wildly disproportionate burden upon a small group. Nonetheless, reducing the total harm remains an important consideration. The tragedy principle promotes the 'societal good', in that the fair distribution of burdens is determined with concern for the society as a whole, rather than concern for the desert of any one individual (as *none* of them deserve the relevant burden).

The tragedy principle has an obvious appeal: where are parties are equally innocent, why should luck determine the proper distribution of losses arising from a potential harm? Yet, the execution of the tragedy principle can sometime be contrary to widely held conceptions of justice. We ordinarily condemn the prejudging of individuals based on traits like ethnicity or religion, even where the prediction is based strictly upon statistical correlations, where it results in the unjust denial of opportunities or distribution of goods, as in racial profiling (Gardner 1998, 168–169; Bonilla-Silva 1997; Boxill 1992, 14; Wasserstrom 1987). Our recent expansion of preventative detention has been limited to clearly defined outsider groups, and even the less controversial practice of quarantine is limited to rare emergency conditions; the societal mainstream has not yet been willing to expose themselves to risk assessment as a condition of their liberty.

It is worth re-examining our grounds for limiting the tragedy principle in these ways, to ask whether they extend to mental illness. At stake in this is our claim to treating people as having equal moral worth. In rejecting the tragedy principle, we grant a tremendous individual right at the expense of the societal good: that we be judged by our own character and choices – traits that respect who we are as *persons* - even where the societal good would be promoted by also judging us by traits that relate to us as subjects, i.e. depersonalised traits. This effectively prioritises the interests of those who are 'morally innocent but dangerous' over those whom they will harm unless detained. By clarifying why our general concern about preventative detention is well-grounded, I hope to better illustrate where the boundaries of our opposition to preventative detention should lie.

3. Prejudice and limitations upon the tragedy principle

In section 1.2 of this paper, I outlined the relationship between mental illness and violence, whereupon the mere fact of mental illness, or even severe mental illness, does very little to indicate the likelihood that a person will commit unlawful violence. That by itself undermines the case for preventative detention of the mentally ill: if the diagnosis of mental illness plays such a minimal role in informing the psychiatric assessment of probable violence, leaving most of the assessment to be informed by more individualised traits such as behaviour, intent, beliefs and character (which would remain dangerous even if explained by something other than mental illness) then it seems unreasonable to take the less relevant factor of mental illness as the deciding trait in determining whether the person is assessed and detained for dangerousness. This is the simplest form of discriminatory injustice: applying a different standard for a group of people on grounds that are not supported by a relevant factual distinction. However, this is an argument for equal treatment, not an argument against preventative detention per se. Once again, I acknowledge that there *may* be some circumstances where a person is so dangerous that we ought to detain them even though they have not yet formed a criminal intent or engaged in criminally dangerous behaviour. Nonetheless, I believe it is worth noting this initial concern regarding preventative detention. Preventative detention, as it is applied to civil commitment, has a *very* broad scope, including *any* risk of even minor interpersonal violence or even property damage. Whilst I do not develop this argument further, I expect that many readers will share my suspicion that if a regime of general preventative detention were imposed universally, we would demand a narrower scope, extreme risk and more rigorous testing.

Wholly aside from the unreasonable distinction in similar levels of dangerousness between persons with and without mental illness, there remains the question of what relevance we ought to place upon the actual relationship between mental illness and potential violence in individual cases. Unjust discrimination isn't limited to false prejudice, but often includes unfair reactions to *true* assertions about a person's characteristics. Again, we routinely condemn discrimination based on race, religion, or ethnicity *even when supported by supported by statistical correlations*, with the leading example being racial profiling (Gardner 1998, 168–169; Bonilla-Silva 1997; Boxill 1992; Wasserstrom 1987). Such statistical data cannot demonstrate that the relevant traits apply to a particular individual, but our condemnation extends to matters to which they are relevant, such as determining the most resource-effective means of policing particular crime. It is not the weakness of the statistical correlation between race and violence that makes racial profiling unjust – after all, were we to cross-study race with enough other factors, such as age, employment, family criminal history and schooling, it is entirely feasible that some sufficient correlation might eventually turn up (if not now, then in the future as actuarial data expands). Rather, we consider the use of such studies to be unethical because it is immoral to judge people by their skin colour instead of matters directly indicative of their persona.

Once we move beyond the obviously unjust phenomenon of false prejudice, the concept of unjust discrimination defies simple categorisation, and I suspect that when we take the tragedy principle into account, the condemnation of such discrimination cannot be adequately explained at the level of individual interests. Early criticism of preventative detention condemned discrimination based on unchangeable traits, on the ground that it unfairly robs people of the ability to control their own destiny

(Underwood 1979). Such criticism has racism and sexism in mind, yet we have little qualms about discriminating on the basis of (largely unchosen) intellect in ways that dominate a person's destiny, and conversely the availability of sex change operations does nothing to justify sexism. Irrespective of whether we can choose sexual orientation, we do choose sexual *activity* – in any event, few advocates of equality would want the wrongfulness of discrimination against gays and lesbians to be contingent upon the empirical claim that sexual orientation is not chosen.

In response, discrimination has been more recently analysed in terms of identity, in the sense of being entitled to make fundamental choices about who one is (Gardner 1998, 170–171). There may be something to this form of criticism, but if so it has little direct bearing upon the issue of mental illness, except perhaps as a factor to be taken into consideration when determining which mental conditions we ought to consider mental illnesses (Edwards 2009, 79–81). Illness (both mental and physical) is something separate to ourselves as persons, ordinarily falling outside the scope of our moral responsibility and autonomy, as something that happens to us, rather than a form of personal action. The very act of labelling something a mental illness precludes our recognition of that condition as a legitimate expression of personal identity. Moreover, this is possible in part because the social stigma attached to serious mental illness, and the debilitating effects of much mental illness, has prevented the conditions' bearers from emphasising their condition as a basis for shared politico-social identity comparable to sexual or racial group identity – contrast, for example, the successful protests brought by gay activists against the listing of homosexuality as mental illness in the Diagnostic and Statistical Manual of Mental Disorders (Graham 2009, 111).

Even aside from this, it is not immediately clear why an interest in making

fundamental choices – to use Gardner's terms, those 'connected with spontaneous self-expression' (1998, 173) – should escape the tragedy principle. As I argued in 'Reasons, autonomy and paternalism' and the last section of 'Mental competence and its limitations', self-authorship is not just a good, but the means by which a person develops and pursues her conception of the good life. Nonetheless, there is a vast philosophical crevice between the pursuit of liberal reasoning where the moral choice is between a person's well-being and her self-authorship, and expanding into broader political liberal theory, in which the individual's interest in self-authorship is in conflict with the interests of other people.. There are few more thoroughly objectifying events than to be made the subject of undeserved violence, and any importance we place upon the detained person's interest in identity will ultimately face comparison to circumstances in which greater harm can be prevented. Even if freedom of self-definition bears the same overriding importance in distributive justice as it does in determining an individual's interests, the self-definition of those with heightened chance of violence would at some point be outweighed by the social obligation to protect the same interest in self-definition possessed by those who would be victims of objectification through violence.

Rather than focusing upon our interest in self-definition, we might instead emphasise the injustice in discriminating on the basis of identity that is implicit in Gardner's (1998) account of discrimination in terms of self-authorship. This would suggest something of a 'does/is' distinction, where it is immoral to discriminate on the basis of choices that are fundamental to an individual's identity, regardless of whether she endorses that identity. For example, a Jewish person's renunciation of her cultural heritage should not justify anti-Semitism against that person, nor should a gay youth's internalisation of societal discrimination justify homophobia. Nonetheless, mental

illness presents a particular difficulty for such a distinction (Edwards 2009, 82-87). There is no objective line to be drawn between the operation of our brains as objects, and those brain processes to which we attribute personhood and character. It is a social and normative distinction, albeit one that ought to be thoroughly informed by the sciences of psychiatry and psychology.

Moral personhood is not a quality that we are objectively capable of to any great degree, but a status that we attribute to others whom we wish to recognise as members of society and participants in social discourse over what is reasonable. Reasoning is a practice carried out between people, where during communication we give and recognise reasons for action 'Reasons, autonomy and paternalism' (also see Hill 1989, 97-98). By acknowledging an individual's participatory membership in society (i.e. personhood), we authenticate her capacity to take part in this giving and recognising of reasons, so that we cannot simply dismiss her attribution of value to goals and values without some appeal to shared values, commitments, duties or authorities. This makes the obligation to respect personhood a politico-social principle, concerning the proper requirements for inclusion as members and participants in society. By judging a person by depersonalised mental illness, we treat her as a subject of society, rather than a participant. Of course, mental illness has an exceptional significance in that it can strip a person of the capacities that we ordinarily insist upon as criteria for personhood. However, preventative detention operates on the basis of risk, not mental incompetence (again, I have no objection to the use of treatment orders to divert ill people out of the punitive justice system in the event that they are mentally incompetent for their crime; though I do not see how this could justify depriving such people of the procedural and evidentiary safeguards available to others accused of similar harm).

An analysis that is restricted to a comparison of individual persons' interests cannot explain why we ought not apply the tragedy principle. To justify a demand for respect even at the cost of some greater harm to others, we need to include a claim to moral *entitlement*, i.e. that those detained have a *right* to be respected as persons despite the loss imposed upon others. Given that neither those to be detained *nor their potential victims* deserve their losses, the entitlement to respect must arise at a societal level.

4. Political disenfranchisement and risk-based detention of the mentally ill

The apparent nobility of the tragedy principle is that it determines fairness by reference to the society as a whole. It appeals to an apparent civic duty to sacrifice our interests for the greater good or, more specifically, the duty of a state charged with protecting its populace to impose such burdens as are necessary to avoid serious harm. In opposing this principle in the context of unjust discrimination, I am neither supporting libertarianism nor criticising the burdening of individuals for the benefit of the group. Instead, the moral challenge brought by the tragedy principle is that of civic inclusion: distinguishing between members of a group sacrificing their interests for the common good, and between *exploitation* of a portion of society by politically stronger components. To make a simple analogy, there is something deeply unjust about defending a state's endorsement of slavery by reference to the benefit to the slave-owners. To talk of the *net* benefit of slavery overlooks the injustice by which the benefit is accrued. More specifically, it treats the slaver-owner and slave as part of the same moral grouping – on the logic of 'net' trade-offs, the interests of both parties are counted together, as one morally homogenous group.

The problem of exploitation questions this assumption of one moral grouping. This is not simply through the idea of 'dirty money', or goods whose worth is tainted by reason of their origins. The core assumption of utilitarianism is that the entire world is one moral grouping, such that any of us are subject to being traded off for a greater net benefit given to someone else we will never meet. The tragedy principle is not a statement of utilitarianism, but it rests on the same assumption of moral unification. The tragedy principle operates where the state has a positive duty to interfere for the sake of the social good, and in pursuing that duty the state presumes that those burdened are a part of the society being protected. If they weren't – for example, if the state was to pursue the social good by stealing from neighbouring societies (say, by invading and exploiting their neighbour's resources), we would need some further moral framework dealing with the justification of interactions *between* societies. Frameworks of this kind are, broadly speaking, a great deal more complex than the tragedy principle, and go to the very heart of basic moral inquiry: how ought we live when our interests seem to conflict with those of our neighbours'. If preventative detention is to be justified by the greater social good, as per the tragedy principle, then it is vital that those detained are part of the same society as those who benefit. That is, it is vital that this is sacrifice *by* the group *for* the group, not the exploitation of one group by another.

I believe that something like an awareness of the difference between exploitation and sacrifice for the sake of the group informs our opposition to sexism and racism. In reality, I strongly doubt that racism or sexism could promote the net good of a society, but our opposition to sexism and racism is not contingent upon that doubt. If confronted by evidence that the benefit conferred upon white men by racial segregation and sexist gender norms outweighed the harm involved, the proper

response would not be to accept oppression, but to point out that this approach to moral reasoning is deeply flawed. The 'net good' is simply irrelevant under circumstances in which the disadvantaged group is excluded from full participatory membership in the broader society, whether through lacking the rights basic to participation in that society (voting and other participation in the process of government, self-ownership, equality before the law, etc), or through policies of detention that actively disrespect their personhood.

An account of legal justice requires justification of the imposition of *authority*. It is not enough that the authority is imposed in order to promote the net societal good. For example, it is simply irrelevant to the injustice of slavery to ask whether or not the benefit to slaveowners outweighed the burden imposed upon those enslaved – oppression of a populace cannot be justified by the benefit conferred upon their oppressors. Aside from any issues of fairness *within* a legal system, the imposition of law itself can be a tool of oppression where those oppressed do not have a fair stake in the legal process. At the very least, legal authority must derive its legitimacy from the society upon which it is imposed. That is, the people subject to the legal authority must have the opportunity to fairly participate as members of the society, such that the laws are the formalised authority of their *own* society rather than a mere assertion of power.

That is not to say we enter into any 'social contract', explicit or implied, giving us a right to have our interests promoted by societal policy. Instead, the attribution of personhood is a fundamental part of our interaction with other humans. Personhood attributes not only the status of moral responsibility, but also the capacity to give and

recognise reasons for action. The 'society' we are concerned with when attributing personhood does not need to be limited to that of a national populace, but rather that collection of people with whom we expect to be capable of engaging in comprehensible debate on a topic (again, personhood is decision-relative), by virtue of recognising the same sources of reason and authority, or at least mutually acknowledging the need to respect each other's contrary sources of reasons ('Reasons, autonomy and paternalism'; also see Hill 1989, 97-98). Some sources of rational authority, however, must be determined at the state level for practical reasons – illness, in particular, demands some shared definition due to the need to establish communal healthcare policies, e.g. for allocation of resources and recognition of illness' role in dismissing personal responsibility. For this reason, i.e. the many policies and terminologies that require the capacity for inclusive discussion that are determined at the level of the state, the state is the effective measure of 'society' for the attribution and recognition of personhood.

Most importantly, the state is the level at which legislative, policing and military power is exercised upon the individual. *One* of the moral failings involved in sexism and racism is the disenfranchisement of a section of the population from participation in the systems that make the exercise of power ethically acceptable in a democratic society. This is not merely a matter of the power to vote – it is the ability to participate in the society itself, benefiting from the society that is protected by the use of power, and doing so as a participant rather than a subject reliant upon the generosity of her social benefactors.

An institution denies people the opportunity to participate when it imposes terms on the basis of things that are beyond their moral responsibility, including their depersonalised traits. Even if a person can change the trait (say, through a sex change as a response to sexism, or medication as a response to marginalisation of the mentally ill), the institution engages with her only as the bearer of that particular trait, not as a morally responsible 'chooser'. As such, her relation to the institution is not that of an agent capable of interaction, but that of an object that is acted upon. The institution acts upon her as something to be governed, imposing terms that are arbitrary with regard to her personhood and character.

This 'acting upon', in the context of societal institutions, is a form of exploitation. That is, it seeks to derive a social benefit from the terms that are imposed, without the participation of those whom the terms are imposed upon. Preventative detention of the 'mentally ill and dangerous' will often improve their lives, and in the last part of this piece I suggest that intervention may often be permissible on paternalistic grounds. Nonetheless, insofar as our purpose is to prevent harm to others, rather than to protect the detainee's own pursuit of projects and commitments, we involve the detainee only as something to be controlled and governed. And insofar as this depersonalisation is imposed upon the detainee, she has no stake in the intervention nor in the security it promotes.

This disenfranchisement is a harm carried out upon the exploited group *as a group* – not just upon that minority who end up being detained. Even when the mentally ill are not immediately detained, the legal permissibility of their detention permeates the

entirety of their engagement with societal institutions – they interact as individuals whose personhood is irrelevant to their liberty, and whose depersonalised traits give reason to interrupt all aspects of their free lives. The wrong is not just the act of detention, but the act of judging a group of people as though they are not people, but mere objects to be analysed in accordance with passive traits instead of person-characterising features such as intent, personal history and the choices they have made. It comes from reducing them, in the eyes of the state, to a collection of illnesses, genders, skin colours, cultures, family histories and education levels. In this social and political context, this objectification undermines the legitimacy of the state's authority. From the moment that one group is excluded from our society through depersonalisation, then at the very least (ignoring the wrong inherent in depersonalisation itself) their own welfare must be protected independently of the 'net' benefit to society. Anything less is wrongful exploitation of the depersonalised group.

The claim to respect, therefore, is a demand to be free from exploitation. The imposition of social authority upon an individual comprises exploitation when that individual is denied fair participation as a member of that society. Such an individual is acted upon for a purpose that he is excluded from partaking in. That is not to say that we *only* owe moral obligations to each other via membership of a shared society; this is not a contract theory of morality. But when participating as members of a society, there is at least the potential (though not always actualised) for social authority structures to have an ethical grounding in our interactions. Without such interaction, our imposition of authority is exploitation: we might impose upon another for our own safety, for someone else's safety or for the pursuit of a moral principle, but in each case we use the targeted person for a purpose that he has no fair stake in.

I have dedicated a lot of space to a series of concepts that I believe should strike most people as relatively simple. I have explained my concerns in this extensive manner because, despite my belief, preventative detention has been almost entirely uncontroversial as a form of civil commitment for a very long time. Indeed, it is nearly impossible to find criticisms of the practice, excluding those who radically reject the entire concept of mental incompetence (Szasz 2008; Szasz 2003), and those who fail to distinguish between the statistical likelihood of violence amongst the mentally ill 'as a whole' and of the smaller proportion who are detained (e.g. Morse 1982). I do not claim to have 'disproved' the moral argument in favour of preventative detention, largely because to the best of my knowledge no substantial argument has ever been published, with virtually all justificatory attention devoted to the more controversial ground of psychiatric *paternalism*. Instead, I hope to have created *doubt* – enough doubt to pass the onus back to those who would support the continuation of such practices – by raising a number of logical incongruities and moral hazards. Firstly, if mental illness plays such a minimal role in determining the likelihood of dangerousness, why should we single out the mentally ill for preventative detention? Secondly, there follows the question of why, if the mentally ill are not exceptionally dangerous compared to others, should we respond to those who are dangerous because of mental illness any differently to those who are equally dangerous for other reasons. Finally, I have sought to elucidate the appeal of preventative detention through the 'tragedy principle', and without claiming that the tragedy principle is never morally authoritative, I have combined social, political and personal concepts of discrimination in order to show that the selective preventative detention of the mentally ill is a form of unjust discrimination in the manner of sexism, racism and religious discrimination. In

the section to come, I discuss one possible outcome of abolishing preventative detention as a ground for civil commitment. Under such an outcome, abolition would not be so radical a reform as it may initially appear, but would deliver tangible justice to those people with mental illness who interact with psychiatric institutions.

5. Paternalistic prevention of violence

The cessation of preventative detention as a ground for civil commitment does not mean that psychiatric paternalism has no business preventing violence. The legal framework for psychiatric paternalism has long recognised that self-harm extends beyond loss of health, to include social harms such as damage to reputation and relationships (e.g. Mental Health Act 1996 (WA) s26(2)). Before leaving this topic, it is worth noting that the traditional informed consent doctrine risks a similar discriminatory injustice through the operation of paternalism based on such social forms of self-harm as that which occurs in preventative detention. I have already discussed at length in 'Mental Competence and its Limitations' how a focus upon mental competence rather than authenticity invites the imposition of the institution's or medical staffs' values upon the patient under the guise of a supposedly objective (but actually at least partially evaluative) standard for mental competence. In the context of alleged self-harm through damaged reputation or relationships, this leads to the risk that a person's competence for decisions regarding social behaviour will be judged in accordance with the person's compliance with the dominant community values, or even just those values held by the individual or institution testing for incompetence. That is, we risk a system wherein the mentally ill face extraordinary restrictions upon their social values, merely by virtue of having a mental illness.

Through Part A of this dissertation, and particular the last two parts of 'Mental competence and its limitations' I have developed an account of patient autonomy that is centred upon self-authorship, rather than freedom or non-intervention. In doing so, I have deliberately moved away from the ideological championing of liberty, to that aspect of liberalism that I believe informs our medical preferences at the level of common-sense judgements: our interest in living our lives in accordance with our own goals and values. This principle can protect the mentally ill from the aforementioned discrimination, in that the significance of any harm (whether to health or to relationships) is determined by the patient's *own* authentic values (except where the length and severity of incapacitation, or absence of possible information about the patient's values, makes meaningful assessment of her authentic values impossible). Moreover, this same goal can provide an ethical underpinning for the prevention of the more damaging forms of 'social' self-harm, and in particular the threat of uncharacteristic violence arising from mental illness.

The motivation for paternalistic assistance in such scenarios should not strike people as bizarre. Even in the absence of mental illness, part of what makes someone a 'good friend' is a willingness to intervene in certain of one's follies, including a willingness to prevent one from acting destructively and out of character when intoxicated. Obviously, there is a limit to the extent to which anyone could be reasonable expected to keep on intervening in such a manner; but nonetheless, it is ordinarily an act of friendship to prevent a friend who is intoxicated or otherwise mentally incapacitated from assaulting another person (assuming that he or she would not be violent if mentally unimpaired). We may be justified in taking such preventative action out of

concern for the potential victim, but that does not explain the full character of the action. It is an action *of friendship*, arising at least in part from one's relationship to the would-be assailant. There is some sense in which we feel we are protecting our friend's interests by intervening. It is difficult to characterise this as conveying some moral benefit, as the friend's state of intoxication and mental incapacity would undermine any claim to moral guidance. Instead, we are paternalistically protecting the friend from a perceived harm. Moreover, social acceptance of such intervention is not entirely dependant upon the risk of criminal punishment. *Obviously* the risk of prosecution is one factor that may make us grateful if our friends prevent us from engaging in harmful idiocy, but a large part is that we wish our own behaviour to be consistent with our goals and values.

This illustrates another point I make in 'Mental competence and its limitations': that autonomy is sometimes a matter of social cooperation rather than freedom from intervention (Meyers, 2005). Similarly, the aspects of self-authorship that matter most to us are very often those that involve our interpersonal relationships and which depend partly upon the way in which we are perceived by others. I argue in 'Beyond Mental Competence' (Edwards 2010, 283) that self-authorship is not wholly determined by our subjective experience of life. Regardless of whether mental illness or sheer moral failure censors violence from your own telling of your personal narrative, the facts that others characterise you as dangerous, that they fear you or have been harmed by you, and that the factual and social telling of your story is one of violence to others, all matter to determining your character. In addition, violence has an exceptional power to define one's relations and character, in comparison to one's other characteristics and history. Violence radically redefines the terms of a social

relationship such that it becomes the dominant feature in the relationship's characterisation. By altering the personal relations that are central to our own identity, roles defined by violence – e.g. 'I can't control my anger', or 'other people fear me' - carry exceptional weight in determining our self-narratives.

It is this sabotaging of self-authorship – the derailing of the person's projects, relationships and commitments – that justifies the state's intervention. There are, perhaps, more tangible ways in which violence restricts one's freedoms, the foremost being the possibility of imprisonment. But if autonomy is to be the grand goal that liberals attribute to it, it must allow different people and cultures to live in the manner they desire; it would be a cruel irony for liberal autonomy to be yet another excuse for imposing our way of life upon others. As such, people must be free (from paternalism) to choose ways of life that lead to restriction of their later freedoms, even eventual imprisonment.[10] By contrast, the sabotaging of self-authorship undercuts the purpose of liberal limitations upon paternalism: such intervention is not seeking to impose some unwanted benefit, or even to impose 'the liberal way of living', but to assist the person in achieving the way of life and degree of independence that he seeks. In particular, I note that if self-authorship is our concern, then state intervention can only be justified if it is a lesser usurping of the person's preferred narrative than the results of non-intervention. This gives us a measure that protects the liberty of the ideological individualist, and requires that intervention be compatible with a person's culture. At the same time, it allows the rest of us – the many people who would rather someone stop them and hospitalise them if necessary, than have their relationships and projects sabotaged through illness-induced violence – the support that our autonomy requires. It

10 Of course, imprisonment itself may be part of the manner in which a person's narrative is derailed.

recognises that for a great many people, autonomy is not achieved alone, but with the support of our families, friends and communities. The nature of mental illness, where intervention may mean hospitalisation and medical treatment, makes the state the appropriate agent of intervention, but the principle is similar: autonomy achieved through support networks that assist our own efforts.

6. Tangible Justice

The moral importance of putting to rest an unjustly discriminatory basis for detention should not be understated. By continuing a policy of detention on the basis of illness and risk, we deny those with mental illness the right to be judged on their actions and intentions as persons, rather than predictions drawn from their depersonalised traits. The rest of us are never required to justify that we are sufficiently 'safe' to live in the general community, regardless of what thuggery we may have demonstrated over the course of our lives. By denying people with mental illness this same right, and we engage in the same kind of bigotry that we have resoundingly rejected in relation to race, religion and gender.

At the same time, there has long been a legal basis for paternalistic protection of mentally ill people from violent conduct that would damage their relationships and reputation, and I have sought to briefly illustrate how such a policy could be made consistent with liberal respect for personal autonomy. This is not just a rebranding of the same practical policies. Preventative detention forces medical staff and psychiatric hospitals to serve conflicting interests, seeking a therapeutic relationship (and holding such a relationship out to the patient) while imposing hospitalisation and treatment

decisions on the basis of others' interests. As an involuntary patient, this means that the patient's medical staff make his medical decisions on his behalf, but not in his interests. Trust in the patient/doctor relationship might sometimes be made impossible by the effects of the patient's illness, but the use of preventative detention does something far more damaging – it makes trust undeserved.

Under a purely paternalistic system, civil commitment can only be justified if (a) the patient will actually benefit from the intervention, and (b) the intervention respects the patient's authentic goals and values. Through the imposition of unwanted medical treatment, civil commitment strips the patient of sovereignty over her own body. By eliminating preventative detention as a ground for civil commitment, we insist that decisions about the patient's body be made for her own benefit, and with respect for her own persona. If there is a case to be made that the patient should be detained for the benefit of society, that case can still be made – but subject to the same evidentiary principles and legal protections available to those who are dangerous for other reasons. A patient's medical treatment, however, would and should be bound to that patient's own interests.

References

Allen, M., 2005. Why Specific Legislation for the Mentally Ill. *Alternative Law Journal*, 30(3), p.103.

Arneson, R.J., 1980. Mill versus paternalism. *Ethics*, 90(4), pp.470-489.

Ashworth, A., 2000. *Sentencing and Criminal Justice* 3rd ed., London: Butterworths.

Bonilla-Silva, E., 1997. Rethinking Racism: towards a structural interpretation. *American Sociological Review*, 62(3), pp.465-480.

Boxill, B.R., 1992. *Blacks and social justice*, Rowman & Littlefield.

Carne, G., 2007. Prevent, detain, control and order? Legislative process and executive outcomes in enacting the Anti-Terrorism Act (No 2) 2005. *Flinders Journal of Law Reform*, pp.17-79.

Corrado, M., 1996a. Punishment and the Wild Beast of Prey: The Problem of Preventative Detention. *Journal of Criminal Law and Criminology*, 86(3), pp.778-814.

Corrado, M., 1996b. Punishment, Quarantine and Preventative Detention. *Criminal Justice Ethics*, 15, pp.3-13.

Crimes (Mental Illness and Unfitness to be Tried) Act 1997 (Vic)

Criminal Procedure Act 2004 (WA)

Criminal Procedure (Insanity and Unfitness to Plead) Act 1991 (UK)

Criminal Procedure Rules 2011 (UK)

Dershowitz, A., 1970. The Law of Dangerousness: Some Fictions about Predictions. *Journal of Legal Education*, 23(1), pp.24-27.

Dworkin, R., 1986. Autonomy and the Demented Self. *The Millbank Quarterly*, 64(Supplement 2), pp.4-16.

Edwards, C., 2010. Beyond Mental Competence. *Journal of Applied Philosophy*, 27(3), pp.273-289.

Edwards, C., 2009. Ethical Decisions in the Classification of Mental Conditions as Mental Illness. *Philosophy, Psychiatry, and Psychology*, 16(1), pp.73-90.

Evidence Act 1906 (WA)

Failer, J.L., 2002. *Who qualifies for rights?: homelessness, mental illness, and civil commitment*, Ithica: Cornell University Press.

Fairall, P. & Lacey, W., 2007. Preventative detention and control orders under federal law: the case for a Bill of Rights. *Melbourne Law Review*, 31(3), pp.1072-1098.

Fazel, S. et al., 2009. Schizophrenia, Substance Abuse, and Violent Crime. *Journal of the American Medical Association*, 301(19), pp.2016 -2023.

Fazel, S. et al., 2010. Bipolar Disorder and Violent Crime: Time at Risk Reanalysis. *Archives of General Psychiatry*, 67(12), pp.931-938.

Federal Bureau of Investigation, 2010. Arrests. *FBI Uniform Crime Reports: Crime in the United States 2010*. Available at: http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/persons-arrested/arrestmain [Accessed January 14, 2012].

Feinberg, 1984. *Harm to Self*, New York: Oxford University Press.

Florida Mental Health Act 1971

Gardner, J., 1998. On the ground of her sex(uality). *Oxford Journal of Legal Studies*, 18, pp.167-187.

Graham, G., 2009. *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*, Taylor & Francis.

Gray, A., 2004. Detaining Future Dangerous Offenders: Dangerous Law. *Deakin Law Review*, 9(1), pp.243-259.

Gray, A., 2005. Standard of Proof, unpredictable behaviour and the high court of Australia's verdict on prevetnative detention laws. *Deakin Law Review*, 10(1), pp.177-207.

Groethe, R., 1977. Overt Dangerous Behaviour as a Constitutional Requirement for Involuntary Civil Commitment of the Mentally Ill. *The University of Chicago Law Review*, 44(3), pp.562-593.

Kable v Director of Public Prosecutions (NSW) [1996] HCA 24; (1996) 189 CLR 51

Kelly, M., 2008. Lock them up - and throw away the key: the preventative detention of sex offenders in the United States and Germany. *Georgetown Journal of International Law*, 39(3), pp.551-572.

Kitai-Sangero, R., 2009. The Limits of Preventative Detention. *McGeorge Law Review*, 40, pp.903-934.

Kleinig, J., 1983. *Paternalism*, Totowa: Rowan and Allenheld.

Lavin, M., 1995. Who Should Be Committable? *Philosophy, Psychiatry, & Psychology*, 2(1), pp.35-47.

Link, B. & Stueve, A., 1994. Psychotic symptoms and the violent/illegal behaviour of mental patients compared to community controls. In J. Monahan & H. Steadman, eds. *Violence and Mental Disorder: Developments in Risk Assessment*. Chicago: Chicago University Press, pp. 136-164.

Lippke, R., 2008. No easy way out: dangerous offenders and preventative detention. *Law and {hilosophy*, 27, pp.383-414.

Mental Health Act 1983 (UK)

Mental Health Act 1983 (SA);

Mental Health Act 1996 (WA)

Mental Health Act 1996 (Vic)

Mental Health Act 2000 (Qld)

Mental Health Act 2007 (NSW)

Meyers, Diana, 'Five faces of selfhood', In *Autonomy and the Challenges to Liberalism*' ed. John Christman and Joel Anderson, Cambridge: Cambridge University

Press: 27-55.

Miranda v State of Arizona, Westover v United States; Vignera v State of New York; State of California v Stewart 384 US 436 (1966)

Montague, P., 1999. Justifying Preventive Detention. *Law and Philosophy*, 18(2), pp.173-185.

Morse, S., 1982. A Preference for Liberty: The Case against Involuntary Commitment of the Mentally Disordered. *California Law Review*, 70(1), pp.54-106.

Morse, S., 1996. An Essay on Preventative Detention. *Boston University Law Review*, 76, pp.113-155.

O'Connor v Donaldson, 422 US 563 (1975)

Papageorgiou, A. et al., 2002. Advance directives for patients compulsorily admitted to hospital with serious mental illness. *The British Journal of Psychiatry*, 181(6), pp.513 - 519.

Reznek, L., 1991. *The Philosophical Defence of Psychiatry*, Routledge.

Robinson, P., 2001. Punishing Dangerousness: cloaking preventative detention as criminal justice. *Harvard Law Review*, 114(5), pp.1429-1456.

Savulescu, J. & Momeyer, R., 1997. Should informed consent be based on rational beliefs? *Journal of Medical Ethics*, 23(5), pp.282-288.

Schoeman, F., 1979. On incapacitating the dangerous. *American Philosophical Quarterly*, 16(1), pp.27-35.

Sentencing Act 1991 (Vic)

Sentencing Act 1995 (WA)

Slovic, P. & Monahan, J., 1995. Probability, danger and coercion: A study of risk perception and decision making in mental health law. *Law and Human Behaviour*, 19(1), pp.49-65.

Swanson, J., 1994. Mental Disorder, Substance Abuse and Community Violence: an

Epidemiological Approach. In H. Steadman, ed. *Violence and Mental Disorder: Developments in Risk Assessment*. Chicago: Chicago University Press, pp. 101-119.

Swanson, J. et al., 1996. Psychotic Symptoms and the risk of violent behaviour in the community. *Criminal Behaviour and Mental Health*, 6(4), pp.309-329.

Szasz, T., 2003. Psychiatry and the control of dangerousness: on the apotropaic function of the term "mental illness." *Journal of Medical Ethics*, 29(4), pp.227-230.

Szasz, T., 2008. *Psychiatry: The Science of Lies*, Syracuse, New York: Syracuse University Press.

Underwood, B., 1979. Law and the Crystal Ball: Predicting behaviour with statistical inference and individualisted judgment. *Yale Law Journal*, 88, pp.1408-1448.

United States v Graydon Earl Comstock, 560 U.S.___(2010)

VanDeVeer, D., 1986. *Paternalistic Intervention: The Moral Bounds on Benevolence*, New Jersey: Princeton University Press.

Ward, M. & Savulescu, J., 2006. Patients who challenge. *Best practice research Clinical anaesthesiology*, 20(4), pp.545-563.

Wasserstrom, R., 1987. Preferential Treatment, Color-Blindness, and the Evils of Racism and Racial Discrimination. *Proceedings and Addresses of the American Philosophical Association*, 61(1), pp.27-42.

Summary of Findings

In adopting the 'series of papers' dissertation format, as permitted by the rules of the UWA Graduate Research School, I have bookended this series with an introduction and summary of findings, to clarify the manner in which the practical and theoretical aims of these papers are connected. The use of distinct papers permits a modular approach, in which I can divide the dissertation's broader aims into independent proofs of each logical step, and then address separately a variety of problems in applied bioethics. Nonetheless, I intend that the papers build towards a greater philosophical program.

I have two broad concerns that I hope to address in this summary. Firstly, I want to take a half step back from the modular conclusions of each piece, so that I may elucidate the practical relevance of these findings as a whole. While this is mostly a work of moral philosophy rather than jurisprudence, it bears significance for law reform (if nothing else, our laws governing involuntary treatment indicate the institutional and state views about patient autonomy). Moreover, an exercise in applying the practical implications of my findings to civil commitment legislation is probably the most transparent means by which the applied implications of these works can be summarily illustrated. As such, I examine their relevance to the civil commitment legislation in Western Australia. I have chosen the Western Australian law as a base because of my past professional experience as a lawyer specialising in representation of the mentally ill in that particular jurisdiction, though it has the further advantage that as a common law jurisdiction it is in many ways representative of involuntary treatment laws in Australia, other Commonthwealth jurisdictions (in particular the UK) and the USA (though, as always, there is considerable diversity

between the numerous US jurisdictions).

Despite the legal significance of my account, I remain concerned primarily with constructing a *moral* case for reform. Rather than engaging in a project of legislative drafting, with its inevitable compromise for the sake of cost control, procedural fairness, legal certainty, dissuading over-litigation and the myriad of further considerations that go into sound law reform, my interest is in discussing the *aims* to which our laws should aspire; identifying the significance of my findings by comparison with the aims that our laws currently manifest. By doing this I hope to summarise the most immediately practical implications of my findings, pointing to those features that (if I am correct) demand substantial alterations to our policies with regard to involuntary psychiatric treatment. That is, I hope to show how the practical implications of my arguments provide an opportunity to rectify significant injustice.

Secondly, there is a need to address the broader relevance of these findings for moral theory. I outlay this dissertation's contribution to broader philosophical trends in the projects of bioethics, philosophy of cognitive science and – most central to my interests – the relationship between moral reasoning and medical science. I suspect that there has been a slow shift in moral thought towards the reclassification of matters that would ordinarily fall within the proper scope of moral deliberation, as matters of scientific judgment. Regardless of whether I am correct in this very broad suspicion, I fear that the structure of orthodox bioethics would encourage this trend (as argued in sections 2 to 4 of 'Mental competence and its limitations'), such that the selective outsourcing of moral deliberation to medical science would be a natural outcome of this way of envisaging the attribution of moral responsibility and autonomy. This would not necessarily be problematic if medical science was capable of providing a

value-free basis for attributing moral agency, but there is no such basis upon which the attribution of illness, competence or personhood can take place. Instead, we risk using medical science as a disguise for the imposition of values other than the patient's, without the moral justification that such imposition should require. My overarching account, then, is a plea within bioethics to acknowledge our full responsibility for societal choices in the attribution of moral responsibility and personhood to other reasoning beings.

Practical implications for law reform

The *Mental Health Act* (WA) 1996 is the legislation governing involuntary psychiatric treatment in the state of Western Australia. Eligibility for involuntary treatment is established by section 26:

> "26. Persons who should be involuntary patients
>
> (1) A person should be an involuntary patient only if —
>
> (a) the person has a mental illness requiring treatment; and
>
> (b) the treatment can be provided through detention in an authorised hospital or through a community treatment order and is required to be so provided in order —
>
> (i) to protect the health or safety of that person or any other person; or
>
> (ii) to protect the person from self-inflicted harm of a kind described in subsection (2); or
>
> (iii) to prevent the person doing serious damage to any property; and

(c) the person has refused or, due to the nature of the mental illness, is unable to consent to the treatment; and

(d) the treatment cannot be adequately provided in a way that would involve less restriction of the freedom of choice and movement of the person than would result from the person being an involuntary patient.

(2) The kinds of self-inflicted harm from which a person may be protected by making the person an involuntary patient are —

(a) serious financial harm; and

(b) lasting or irreparable harm to any important personal relationship resulting from damage to the reputation of the person among those with whom the person has such relationships; and

(c) serious damage to the reputation of the person."

As mentioned in the opening sections of 'Mental competence and its limitations' and 'Suicide prevention and the limits of patient autonomy', an initial reading of this section gives the appearance of a transparently single-minded purpose: the prevention of harm arising from mental illness, where that harm is defined by a cumulative combination of medical, societal and personal standards (see the references to harm to reputation, finance and relationships, in addition to physical harm). Again, as discussed in those same paper, there is considerable room for argument that the aim of the legislation is something closer to that espoused by orthodox bioethics: that same concern for preventing harm to the patient and others, but limited (at least with regard to paternalistic treatment) to those patients whose illness renders them mentally incompetent.

Mental illness is only loosely defined in this Act:

> "4. Mental illness, defined
>
> (1) For the purposes of this Act a person has a mental illness if the person suffers from a disturbance of thought, mood, volition, perception, orientation or memory that impairs judgment or behaviour to a significant extent.
>
> (2) However a person does not have a mental illness by reason only of one or more of the following, that is, that the person —
>
> (a) holds, or refuses to hold, a particular religious, philosophical, or political belief or opinion;
>
> (b) is sexually promiscuous, or has a particular sexual preference;
>
> (c) engages in immoral or indecent conduct;
>
> (d) has an intellectual disability;
>
> (e) takes drugs or alcohol;
>
> (f) demonstrates anti-social behaviour." (section 4, *Mental Health Act* (WA) 1996).

I note in the opening section of 'Mental competence and its limitations' that similar lack of specificity is found in equivalent legislation from comparable jurisdictions such as *Mental Health Act* (UK) 1983 and *Florida Mental Health Act* 1971. Aside for making room for the incorporation of something akin to a mental incompetence criterion into the legislative test for involuntary treatment, it reflects an outsourcing to the psychiatric fraternity the job of identifying which mental conditions ought deprive an individual of moral agency, with some highly notable exceptions listed in s(4)(2).

This is a suitable point at which to outline my own summary of aims that civil commitment legislation should aspire to. These are not comprehensive, but reflect only my findings in the course of this dissertation:

1. To establish a moral concept of mental illness constructed upon a liberal concept of personhood, which provides a just basis for stripping ill individuals of moral agency. This concept of illness should be informed by medical science with regard to the operation of brain processes, the effects of mental conditions upon those processes, and the relationship between these conditions and the exercise of moral agency.
2. To respect and protect the personal integrity and authenticity of patients, from the impositions of illness and unjust intervention alike.
3. To identify the absolute limits of liberal personhood and provide psychiatric intervention to treat and prevent harms beyond those choices which can form expressions of personal identity.
4. To acknowledge, respect and protect the interests of both permanently impaired patients and the persons they once were, by resolving conflict between these two sets of interests in a manner that diminishes neither of them.
5. To abolish the use of involuntary psychiatric treatment for any purpose other than the protection of the patient being treated. Simultaneously, to acknowledge that the social nature of our personal identities and autonomy indicates that the imposition by mental illness of violent behaviour is often a serious usurping of personal integrity, and thus a harm warranting paternalistic intervention.

The first aim is the subject of the first substantive paper in this dissertation, ‘Ethical decisions in the classification of mental conditions as mental illnesses’ (Edwards

2009a). I stated at the end of that paper that I do not intend to 'steal expertise away from psychiatry' (2009a, 87). Jerome Wakefield charges that my account does exactly that (J. Wakefield 2009, 94). I responded to this in 'Changing functions, moral responsibility and mental illness' (Edwards 2009b), attached as a schedule to this dissertation, but Wakefield's concern is essentially a product of confusion over the multiple meanings of 'illness'. My concern is strictly with the moral concept: that which justifies our practice of not only depriving the patient of autonomy, but also freeing her of responsibility for the symptoms of the illness, morally justifying the separation of the illness from her own persona. For Wakefield (2009, 92), pedophilia is a form of illness, and perhaps this is correct insofar as it is a condition that warrants psychiatric treatment. But it in no way occupies the moral status of illness: we apply extreme moral, and even criminal, responsibility for the condition of paedophilia and view it as a defining feature of the person's character.

In defining this *moral* concept of mental illness, medical science has never held a monopoly; if anything the dominant authorities in defining illness as a moral concept have been those responsible for encoding and enforcing the boundaries of social morality in the form of *law*, i.e. the legislators and courts responsible for deciding whether a mental condition should absolve individuals of their crimes, or render patients vulnerable to involuntary hospitalisation. That the legislature and courts now defer to psychiatric expertise in carrying out this duty is clearly an improvement insofar as it results in policy based on a scientific comprehension of mental processes, the impact of illness upon these processes, and the medical treatments available. Nothing in my account denies that we should continue to rely on such expertise. In identifying the normative aspects of 'proper function', I call for societal examination of the values by which mental dysfunction and illness are defined. However, this can

only be productive if such examination is informed by medical science. Most likely, as the segment of the population best educated about the relevant science, psychiatrists will still be called upon to define the moral concept of mental illness, subject to legislative exceptions of the kind already present in s4(2) of the *Mental Health Act* (WA) 1996. However, rather than viewing such legislative intervention as populist intrusions upon science, they should be encouraged as part of the process by which mental illness (in the moral sense) is kept consistent with societal concepts of personhood and moral responsibility.

The second of the above aims is stated briefly, but is the ultimate outcome of the account I construct over the course of Part A of this dissertation. I expressly construe this account as one that takes place *within* the medical tradition of liberalism with regard to paternalism, embodied in law and general medical practice through the informed consent doctrine, and near universal within theoretical bioethics. I spend the initial sections of 'Mental competence and its limitations' explaining that orthodox bioethics renders psychiatric paternalism a special case of medical paternalism, subject ultimately to the informed consent doctrine and respect for mental competence as a set of qualities that qualifies a person for unfettered choice to accept or refuse medical treatment. The informed consent doctrine endeavours to provide an unbiased method for identifying those people whose judgment is so greatly impaired that they should be stripped of both moral responsibility and autonomy – i.e. it seeks to separate the *content* of the patient's choice from the *justifications* for overriding that choice, such that the patient's authority is determined solely by her information and mental capacities and not the eccentricity of her decision or personal values. The informed consent doctrine, and the concept of mental competence that it turns upon, not only fails to achieve this aim, but risks entrenching at an institutional level the very bias it

seeks to combat. It enables the imposition of the institution's or doctors' values in a manner that is impossible to prove in individual cases, but which nonetheless appears broadly embedded in psychiatric practice.

This is not a failing in the intent or expertise of the psychiatric staff tasked with assessing mental competence, but a consequence of our asking something of them that is impossible. There is *no* objective and value-free set of qualities that qualify us for moral agency, against which we can measure mental competence. Moral responsibility and autonomy are not qualities that we are objectively capable of to any great degree, but statuses that we attribute to each other in recognition that we share certain qualities that we deem to provide an appropriate basis for *attributing* personhood and *demanding* moral responsibility. The standard of mental competence, like mental illness itself, is necessarily an evaluative concept. This renders any protection from paternalism a matter of the values upon which our standards of mental competence are based. If patient autonomy is to have moral meaning in this context, therefore, it must not refer merely to the ability to qualify for independence over individual choices. Patient autonomy must include the means to develop and pursue one's own conception of the good, through self-authorship of a narratively intelligible and authentic personal identity. Under this account then, we respect a patient's autonomy by ensuring that the patient's treatment is governed by values that permit meaningful self-authorship; i.e. that the patient's treatment is governed by her own authentic goals and values. Foremost to this concept of autonomy is not the scope of our freedoms, but the nature of our relationships. Instead of seeking to preserve a right to qualify for independent choice, it is measured against its ability to pursue the patient's own goals and values.

This requires an account of personal identity that enables the conceptual separation of *in*authentic values, from the patient's authentic identity. I provide this account in 'Beyond mental competence' (Edwards 2010) while expanding upon my argument that the justification of psychiatric paternalism cannot be adequately addressed in terms of mental incompetence alone.

Finally, I argue in 'Suicide prevention and the limits of patient autonomy' that this liberal personhood must be finite in scope, defined by positive liberties rather than selective non-intervention, if personal autonomy is to have the moral significance that I and other bioethicists espouse. The value that we invest in personal authenticity is grounded in the capacity for individual persons to imbue their lives and their environment with value. Crucially, many of us strive to imbue our lives with value that extends *beyond* our personal existence, seeking to embody the value of absolutes through our adoption of moral and ideological principles, even at cost to our own well-being, and sometimes even at cost to our lives. The pursuit of this transcendental value is what gives such moral weight to the notion of 'dying with dignity', and we are justified in separating such deaths from those motivated by the denial of self-worth. Liberal personhood only warrants our social authentication when it is self-affirming (though not necessarily *life*-affirming), and psychiatric institutions act appropriately by intervening in self-harm based on the denial of self-worth.

The fourth aim summarises the practical findings of 'Respect for other selves' (Edwards 2011). The immediate concern of that paper was the authority of advance directives in circumstances where the interests of a person who makes an advance directive come into gross conflict with those of the deeply mentally impaired patient whom she becomes. This ordinarily arises where the person has directed the cessation

of life-extending or life-saving treatment, but theoretically it could include the demand that the patient be kept in conditions of great suffering or indignity (perhaps through a religious belief in facing suffering without pain relief, or an eccentricity demanding the continuation of practices that would otherwise be considered demeaning). I concur that advance directives may convey a crucial personal interest, but argue that the worth of the patient's current experiential interests has been widely underestimated. The two require reconciliation on a case by case basis that is at odds with the blanket legal authority that some advocates of advance directives demand of such documents. I suggest in the article the use of substitute decision-makers with concern for both sets of interests.

The last aim is the most direct, in that it is the only possible outcome of my final paper 'Problems with the easy justification'. To strip the mentally ill of the rights and protections given to other legal defendants, subjecting them to examination of dangerousness as individuals and through actuarial study, is not supported by statistical studies of violence, and is morally abominable independently of the statistical insignificance of mental illness. It may be that sufficiently dangerous individuals, regardless of illness, income or other characteristics, could warrant preventative detention. Of course, I strongly suspect that we would insist upon a far more rigorous system of checks and a far greater standard of dangerousness than that upon which we condemn the mentally ill to detention for the protection of others. Nonetheless, to pick out the mentally ill as a group and expose them to non-paternalistic preventative detention is a social and political injustice of a similar form to that of racism and sexism. On the flipside of this criticism, my account of the ethics of psychiatric paternalism encourages a permissive approach to *paternalistic* prevention of violent behaviour. By making patient authenticity and integrity central to respect for patient

autonomy, we ought to recognise that illness-induced violence, with its risk of catastrophic effect upon the close relationships that are central to personal identity, can irretrievably shatter an individual's opportunity to construct an identity and way of life that is consistent with her conception of the good. Involuntary psychiatric hospitalisation and treatment should never be used for any purpose other than the benefit of the person being treated; yet there is much common ground between detention for the protection of others and paternalistic protection of personal authenticity, and prevention of illness-induced violence may well be an appropriate aspect of paternalistic involuntary treatment.

Social morality and medical science

The central outcome of this dissertation is the shift from an objective to an evaluative account of and the justifications for psychiatric paternalism. To summarise the most vital finding from Part A of this dissertation: there *is* no objective standard of mental competence, and thus the attribution of moral agency (both responsibility and autonomy) is a moral and social decision. Moral agency in its current form is not an objective fact that is innate to human existence, and it would have been at least conceptually possible (if not historically plausible) for the role of our current liberal conception of personhood to be fulfilled by religious or feudal conceptions of personhood. 'Liberal personhood' in this sense is, as I explain in 'Ethical decisions in the classification of mental conditions as mental illness' (Edwards 2009a, 78), nothing to do with political liberalism or any claim to individual rights over the public good. It refers solely to our mutual recognition of each other as beings capable of choice over *who we are*, coupled with responsibility for those same choices.

Whether through biological imperative or cultural structures, we *decide* to interact with each other through social structures that attribute to us moral agency. Again, as explained in 'Mental competence and its limitations', this is not an objective implication of human capacities, but something we attribute to each other *despite* our extensive limitations to reasoning, judgment and mood control. There is a moral tension, therefore, between the standards that we insist upon as the minimal criteria for moral agency, and our aim of social inclusiveness.

It seems almost inevitable that our knowledge of the brain, our mental processes and the limitations upon our judgment and mood control will grow tremendously as medical science in the areas of psychiatry and neurology progresses. This is not a work of psychiatric scepticism – indeed there are points in this thesis where I provide a more secure defence for broad psychiatric intervention than that of orthodox bioethics – and this improvement in medical science is likely to be a tremendous boon. Yet, as I identify in Edwards (2009a, 76), our identification of illness (mental and physical) has more to do with the needs we impose as persons than any evolutionary dysfunction, and so even evolutionarily 'natural' limitations to judgment may well become medical conditions for which people legitimately seek psychiatric treatment. After all, if a condition causes someone any amount of frustration or suffering, and a doctor can alleviate that suffering, it would be churlish to argue that the illness and treatment was somehow illegitimate because of an evolutionary normality far removed from the needs of modern society.

Nonetheless, it is vital that we recognise that the medicalisation of such conditions says absolutely nothing, in itself, about whether a patient should be excluded from our 'society of persons' (whether fully, or simply in relation to the symptoms of the

illness). Medical science can *inform* us of the impact of these conditions upon the patient's ability to develop and pursue her authentic personal identity, but the mere identification of a mental illness (in this medical sense) should have no direct moral repercussions. As improvements to medical science lead to an ever-expanding list of mental conditions recognised by medical science as affecting our judgment, this should not directly translate into a similar expansion of conditions brought before the courts as evidence of moral excuse or mitigation. Every time that we recognise a mental condition as an illness in the moral sense, we narrow bit by bit the society of persons who we recognise as equals in the recognition of reasons and the bearing of moral responsibility. A concept of personal autonomy defined by independence rather than personal authenticity would rapidly become an 'autonomy of the gaps': squeezed into the ever-shrinking gap in our knowledge about the mind's causal processes. I fear that the concept of mental competence, and with it respect for personhood, would soon follow, as we become increasingly aware of a possibly enormous list of mental failings that we have hitherto viewed as part of ordinary (and morally culpable) lapses of reason. That is not to say that such discoveries can never result in a morally just revision of our moral agency. It may be that some of our current mental limitations, and widely shared deficiencies, become so widely and easily treatable, with so little consequent social friction, that by tightening the criteria for personhood we do not narrow the society of persons, but instead improve upon what we can justifiably take to be normal function. The vital component of this is that it is not an automatic consequence of psychiatric discovery. This is a matter of moral judgment, not just scientific expertise.

Perhaps most importantly, in 'Problems with the easy justification' I make the case that the danger of encroachment of medical science upon morality takes its ugliest

form when used to exclude individuals from the society of persons in a manner that justifies non-paternalistic detention. Just as rational imperfection is a part of normal humanity, so is risk to others, both actuarial and personal. Of all the consequences of conflating the extension of psychiatric expertise with reduction of the scope of moral agency permitted in our society, this is the most dangerous; that of depriving people of the rights associated with personhood until they may be used as a tool for the greater good, with their detention a means towards the ends of a society from which they are excluded. The consequence of such policies reflects equally poorly upon societal morality and those who would hide such social and political disenfranchisement under the guise of medical science.

As with physical competency, the limits of liberal personhood are not infinite. If we were to insist upon applying full moral responsibility upon those suffering serious delusions or psychotic states, the holding of moral responsibility would become unworkably arbitrary and hostile to the functioning of any sustainable society. Further, as I identify in 'Suicide prevention and the limits of mental competence', it will not do to simply pick out the liberal model of personhood as one of many possible ways of interacting with each other. Even allowing that there may be multiple viable methods of social interaction, liberal personhood must be capable of the moral worth that we attribute to others on its basis. To hold such worth, it must be a vehicle for the attribution of value upon the world and ourselves.

Yet, as I emphasis throughout 'Respect for other selves' (Edwards 2011), the moral authority of liberal personhood is limited by our duties to non-persons, including the permanently mentally impaired. In an age where we recognise that non-persons, in the form of non-humans, can have serious moral worth, we cannot totally strip a human of

their moral claim not to be tortured, or to be kept in demeaning conditions simply because she is no longer part of our society of persons. Moreover, even where the potential loss is merely one of the potential happiness from killing an otherwise contented patient, we must nonetheless balance the interests of the pre-impairment patient with her current experiential interests – and the assumption of the universal authority of person-related interests over post-personhood interests is every bit as morally distasteful as if the post-person retained the verbal capability to plead her (intellectually limited) desire for continued existence.

In summary, this dissertation has argued for the transformation of the bioethics on psychiatric paternalism from objective to evaluative. In doing so, I have not sought to pursue some radical new direction in medical ethics, but to protect the authority of moral conscience from q flawed reliance on medical science and the equating of the *medical* concept of mental illness with the *moral* concepts of illness and incompetence. As is inevitable in such a work, I do not provide answers to many of the moral questions I raise: for example, I do not fully define the limits of liberal personhood, nor do I provide a definitive statement of the appropriate standard for mental competence. In these cases, like much of those addressed in this thesis, the onus of argument has been to show that these are moral questions at all, rather than matters in which societal morality must bend entirely to medical consensus. These further moral questions are not necessarily matters best determined by academic philosophy. I do not hope for councils of moral philosophers to supervise the growth of categories of illness and dysfunction. Instead, if I was to picture an ideal, it would be more akin to that of psychiatrists and courts seeking greater collaboration with political and social interest groups in reaching a mutual understanding of mental illness, competence and authenticity; one in which we recognise psychiatry as the science upon which the

moral dimensions of illness are investigated, rather than an authority instructing society about the standards for illness (as a moral concept) and personhood as statements of medical fact.

References

Beauchamp, Tom, and James Childress. 2009. *Principles of Biomedical Ethics*. 6th ed. New York: Oxford University Press.

Boorse, Christopher. 1997. 'A Rebuttal on Health'. In *What Is Disease?*, ed. James Humber and Robert Almeder, 1-134. Totowa, New Jersey: Humana Press.

Dresser. 1984. 'Bound to Treatment: The Ulysees Contract'. *The Hastings Centre Report* 14 (3): 13-16.

———. 1995. 'Dworkin on Dementia: Elegant Theory, Questionable Policy'. *Hastings Centre Report* 25 (6): 32-38.

Dworkin, Ronald, Thomas Nagel, Robert Nozick, John Rawls, Thomas Scanlon, and Judith Jarvis Thomson. 2007. 'The Philosopher's Brief'. In *Bioethics: Introduction to History, Method and Practice*, ed. Nancy Jecker, Albert Jonson, and Robert Pearlman. 2nd ed. Sudbury, Mass.: Jones & Bartlett Publishers.

Edwards, Craig. 2009a. 'Ethical Decisions in the Classification of Mental Conditions as Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 73-90.

———. 2009b. 'Changing Functions, Moral Responsibility, and Mental Illness'. *Philosophy, Psychiatry, and Psychology* 16 (1): 105-107.

———. 2010. 'Beyond Mental Competence'. *Journal of Applied Philosophy* 27 (3): 273-289. doi:10.1111/j.1468-5930.2010.00491.x.

———. 2011. 'Respect for Other Selves'. *Kennedy Institute of Ethics Journal* 21 (4): 349-378.

Florida Mental Health Act 1971

Grisso, Thomas, and Paul Appelbaum. 1995. 'The MacArthur Treatment Competence Study'. *Law and Human Behaviour* 19 (2): 105-126.

Mental Health Act (UK) 1983

Mental Health Act (WA) 1996

Robertson. 1991. 'Second Thoughts on Living Wills'. *Hastings Centre Report* 21 (6): 6-8.

Wakefield. 1992. 'The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values' 47 (3): 373-388.

Wakefield, Jerome. 2009. 'Mental Disorder and Moral Responsibility: Disorders of Personhood as Harmful Dysfunctions, With Special Reference to Alcoholism'. *Philosophy, Psychiatry, and Psychology* 16 (1): 91-99.

Wakefield, Jerome C. 2006. 'What Makes a Mental Disorder Mental?' *Philosophy, Psychiatry, and Psychology* 13 (2): 123-131.

Schedule A - Changing functions, moral responsibility and mental illness

Abstract: This was a paper arising from the material that went into production of the dissertation. I was fortunate enough to have my first paper, 'Ethical decisions in the classification of mental conditions as mental illnesses' subjected to commentaries by other philosophers, most notably Jerome Wakefield, whose has done much to develop the dominant account of mental illness, which I disagree with in the above paper. I felt that Wakefield's comments arose primarily due to his misunderstanding my own account (admittedly, as my first published piece, I was not then as clear as I may have been if authoring the same piece today) , and especially a confusion as to the nature of illness that I was concerned with. In the following piece I explain what I mean by a 'moral' concept of mental illness and (to my mind) significantly clarify my original work. Nonetheless, the piece adds nothing in terms of actual substance to my original account, and for that reason I attach it only as a schedule to the dissertation

Changing functions, moral responsibility and mental illness

I thank both Wakefield and Tomasini for their illuminating comments. Both commentaries are thought-provoking and warrant a full response. However, as always, space is limited and I must make the all too predictable apology for not addressing both commentaries in full. Wakefield's contribution more directly engages with, and challenges, my claims, and so I will focus upon addressing his concerns.

Clarifications

Wakefield correctly notes that I am interested in how mental illness (and illness generally) operates to remove moral responsibility for one's actions. Despite this, it seems at several points that he and I are talking about two different, albeit related, concepts of mental illness. 'Mental illness' is used to both explain and excuse certain behaviours, but these two functions, explanatory and exculpable, do not always operate simultaneously. As we discover biological explanations for apparently immoral behaviours, the question arises as to whether the explanation should *remove* the immorality and liability for punishment. The 'sick role' can be divided into two aspects. One is the role of deserving and warranting medical treatment. The other is the negation of moral responsibility. Wakefield's own example, paedophilia, involves the first sense of the sick role but not the second. Wakefield's use of this example is unusual, as it deftly highlights the limitation of his HD account. Paedophilia is a mental disorder in the sense of warranting treatment, and is a mental illness under Wakefield's model – but it is not something that negates moral responsibility. There is a sense of the term 'mental illness' – an important and, as I argue, an objective sense – that is relevant to establishing one's moral responsibility for such behaviours.

In addressing my account of mental illness as though it contained two distinct answers, 'mental illness as dysfunction of rational agency, or 'mental illness as moral judgment', it is unclear whether Wakefield takes them to be conjunctive or alternative. Hence it is worth reiterating in point form the structure of my account:

1. Illness involves dysfunction;

2. The 'proper function' of a bodily or psychological mechanism is determined by the purpose that we impose upon that mechanism (within the constraints of our biological needs);

3. Dysfunction of rational agency is necessary, but not sufficient, for 'mental illness' in the normative sense relevant to negating moral responsibility;

4. The distinction between 'mental illness' (in the above sense) and the many dysfunctions of rational agency for which we are morally responsible, turns upon (objective) normative facts rather than features internal to the dysfunctions themselves.

Functions and Value

Wakefield misses the point when he criticises my comment that while bodily mechanisms are fairly uniformly valued, whereas there is substantial historical and cross-cultural variation in the way people value personhood. He notes – accurately, but irrelevantly – that people almost always place great value upon their capacity for rational agency, probably more so than many aspects of their physical health. My comment, taken in context, has nothing to do with *how much* people value personhood, but has everything to do with what kinds of purposes people want their mental and

physical processes to achieve. A heart has the same function for a Londoner today as it did for a Japanese peasant 1000 years ago, and whilst it is possible for us to want it to achieve other purposes, any culture that *seriously* imputes a different set of vital functions upon the heart is likely to be rather disappointed. By contrast, our species has experienced vast change in the demands we make of our rational and intellectual capacities – dysfunctions like dyslexia and alcoholism were irrelevant prior to the availability of written communication and alcohol.

Wakefield's HD account is unable to deal with the relevance of such change. The link between evolutionary function and illness is tenuous in any event. Evolutionary explanations of specific mechanisms are often untestable, simply positing some evolutionary purpose that happens to fit our understanding of what the mechanism does. If we were to discover that alcoholism evolved as a means of increasing birth-rates, or (perhaps more plausibly) that anxiety disorders occur among those for whom the protective evolutionary function of the relevant brain and psychological processes actually work more *efficiently* – ought we then change our view that these are illnesses? What would matter more, the mechanism's correctly fulfilling the purpose of protecting our ancestors from predators and hostile humans, or it marring the goal of social interaction that we attribute to our capacities for rational agency? Or consider the impact of technological implantations such as artificial limbs, pacemakers and so forth. Currently, these are the exception to normal functioning. But again, that is a contingent fact - whether as an improbable thought experiment, or a possible future scenario, we can envisage a scenario where the function we impute upon bodily mechanisms changes due to new needs arising as a consequence of medical technology. To insist that someone suffering the rejection of an artificial limb or other medical device is healthy, simply because the bodily mechanisms are meeting their

evolutionary purposes, would be bizarre.

Values and dysfunctions

Wakefield suggests that his Harmful Dysfunction ('HD') model addresses the question of why some mental conditions justify the sick role and others do not. To the contrary, his own example of paedophilia demonstrates that – with regards to the negation of moral responsibility – that is not the case. Whether we adopt an evolutionary account of dysfunction or my own account, we are still left with the question of why some dysfunctions negate responsibility and others don't. Wakefield's model gives us something approximating:

1. [dysfunction] – body's mechanism not functioning;
2. [harm] – will cause social isolation, etc;
3. [ordinary ethical theory] – value of life, happiness etc;
4. Therefore [normative outcome] – treatment is warranted.

The 'harm' criterion has moral relevance, and hence we can derive the normative conclusion. But the 'harm' criterion has no normative relevance to the negation of moral responsibility – which is the normative conclusion that I am interested in. Applied to that issue, we have (on either Wakefield's or my account):

1. [dysfunction]
2. [harm]
3. [ethical theory], then
4. [normative outcome] *either* (a) moral responsibility is negated *or* (b) moral responsibility isn't negated.

So what gives us 4(a) rather than 4(b)? The harmful dysfunction criteria tell us nothing – paedophilia meets both criteria, without negating moral responsibility. Hence a further explanation is required.

Also contrary to Wakefield's suggestion, there is no peculiar difficulty in identifying impairments of rational agency without reference to evolutionary dysfunction, as in explaining why agoraphobia is not simply an exercise of rational agency in response to stimuli. I am surprised that Wakefield doesn't refer to the enormous wealth of philosophy dealing with precisely the question of what differentiates compulsion from voluntary agency (e.g. Frankfurt 1998, 11-25, 159-176; 1999, 129-141; G Dworkin 1989; Christman 2005). One popular account is that the person doesn't endorse the desire to avoid open spaces, and does have an endorsed higher-order desire to not avoid them (Frankfurt 1998, 11-25). Regardless of whether one accepts that particular explanation, philosophy is a very long way from having to fall upon an evolutionary account in order to explain the distinction.

With his own account making no distinction between dysfunctions that negate responsibility and those that don't (paedophilia being clearly among them), Wakefield compounds this problem by noting that alcoholism also fits his HD model. The problem is that most of us are, for good reason, not yet willing to remove moral responsibility for alcoholism or consequent behaviours. These examples show that the attribution of the label 'mental illness' *in the sense of moral negation* is not simply a matter of harmful dysfunction. However, the alternative explanation is not, as Wakefield suggests, a subjective one that excludes factual diagnosis. The truth of a moral claim may depend partially upon facts about a culture's practices and beliefs, but that doesn't prevent moral claims from being objectively true or false. Moral

philosophy has never fully succumbed to the cultural relativism of mid 20th century sociology, and the ongoing relevance of Kantian and Parfitian accounts show that I am far from alone in this view.

Lastly, Wakefield asks what relevance psychiatric expertise has under my model. I suggested that the negation of moral responsibility turns upon a matrix of factors involving the fairness of attributing moral responsibility and the effects that withholding responsibility will have upon the broader ethical system. To answer these questions, we need to know the details of the dysfunction itself! The factual element of diagnosis is as present on my account as in Wakefield's. Even the normative aspects of my account of proper function do not affect the role of psychiatry – whether proper function is determined by evolutionary purpose or current purpose, we still need medical expertise to determine how that purpose is being impaired. The only difference is that whilst Wakefield lumps paedophilia, alcoholism and schizophrenia in together as all falling within the HD model, I ask the further question of why we negate responsibility in some cases of the third condition, but not the other two.

References

Christman, 2005. 'Autonomy, Self-Knowledge, and Liberal Legitimacy' in *Autonomy and the Challenges to Liberalism*, ed Christman and Anderson, 330-357. UK: Cambridge University Press.

G Dworkin, 1989. 'The Concept of Autonomy' in *The Inner Citadel: Essays on Individual Autonomy*, ed Christman, 54-62. UK: Oxford University Press.

Frankfurt, 1988. *The importance of what we care about: Philosophical Essays*, UK: Cambridge University Press.

Frankfurt, 1999. *Necessity, Volition, and Love*, UK: Cambridge University Press.