The University Of Sheffield.

Understanding the Ethics of Changing Moral Dispositions

Brendan Kelters

The University of Sheffield

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

2022

Understanding the Ethics of Changing Moral Dispositions

Brendan Kelters

Department of Philosophy

June 2022

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Acknowledgements

I would like to thank the many individuals and organisations who helped in the preparation of this thesis.

First amongst these is my supervisor Yonatan Shemmer, whose wide knowledge, razor-sharp mind, and generosity with time have been my constant aids throughout this project. Much of what is now before you was born from some suggestion or critique offered by Yonatan over these past five years and his contribution to my efforts – and indeed my development as a thinker – cannot be overstated. Also supporting me throughout was James 'Jimmy' Lenman, whose wisdom was always there to show me the way forward – or sideways, or back and around – whenever it was needed. I count myself honoured to have been able to benefit so consistently from Jimmy's experience, perspective and kindness while preparing this thesis. Taken together, I couldn't have hoped for better supervisors nor will I ever be able to fully express my gratitude to them.

In addition to my supervisors' efforts this thesis has benefitted enormously from the input of a number of other denizens of the University of Sheffield's Philosophy Department. Amongst staff I owe particular thanks to Jenny Saul for many extremely helpful discussions of my views and material. I also must thank Robert 'Bob' Stern for enlightening me about some important nuances of Kantian views on my subject. Over the years this thesis has also benefitted from various comments of James Lewis, William Hornett, Mathew Cull, Lewis Brooks, Graham Bex-Priestly, William Morgan, Kayleigh Doherty, Ahmad Fattah, Rosa Vince and Dunja Begovic. Beyond the wonderful community at Sheffield, furthermore, I must thank David Archard for introducing me to a perspective on moral psychology which continues to animate my thought. I must also thank reviewers Connie Rosati and Jules Holroyd for their time and careful critique, which has greatly benefitted the project.

Aside from this already remarkable collection of interlocutors I owe further and special thanks to my long-suffering family for their patient engagement with the constant barrage of thought-experimentation and existential interrogation I unleashed upon them while preparing this thesis. To my mother Camilla, my father Seamus, my brother Michael and my grandmother Donna I am in this – as in all things – forever grateful.

While all these many and wonderful people have helped make this thesis what it is all errors are, of course, my own.

Finally, and besides those I've already mentioned, I owe a debt of gratitude to certain people and institutions who have provided my project with invaluable practical support. The White Rose College of the Arts and Humanities not only financed my studies but – particularly in the person of Manager Caryn Douglas – more than once rescued me from the effects of my administrative ineptitude. Within the University of Sheffield's Philosophy Department, the excellent office staff – notably Joanne Renshaw and Patrizia Baldi – have fulfilled much the same role. I similarly owe special thanks to Tom Brocket and his colleagues at the Department for Digital, Culture, Media and Sport for making my stay there in 2019 an enjoyable and interesting one. I will forever remember the remarkable curiosity and conscientious engagement with my work I encountered during that time.

Abstract

People have moral dispositions, parts of their moral psychologies that determine how they act in moral contexts. These dispositions are subject to influence and change, ever more so as we discover more about them and what we can do to effect change in them. This forces us to consider which things ought to impact these dispositions, which things are legitimate influences upon them. I argue that – other considerations aside – perceived relevant reasons ought to influence these moral dispositions. Ceteris paribus, it's always better that such reasons influence moral dispositions, yet better when more do, worse when none do yet dispositions still change. I show this to be the case by analysis of thought experiments which hold other variables constant and propose that this is the case because we demand that behaviour properly connects with and responds to the reasons agents have. I argue that this understanding, captured in what I call a 'rationalizability requirement' on moral-dispositional change, has important implications. It refutes what I call 'instrumentalism' about moral-dispositional influencing; the view that the only thing that matters in evaluating such influencing is results (broadly construed). Methods, it turns out, matter. It also helps us to evaluate techniques used in advertising, brainwashing, moral bioenhancement and nudging. It does all these things, I contend, while garnering additional support from those who value freedom, construed as autonomy or as interpersonal non-domination, and thereby unifying (at least in practice) influential extant perspectives on the evaluation of moral-dispositional influencings into a single standard.

Table of Contents

Chapter 1: Introduction and Summary

(1.1) Why seek an Ethics of Moral-dispositional Change?

We have morals. These morals can change. We – in dealing with ourselves and others – can have the power to exercise some control over this change. How should we exercise this control?

This question is an ethical one. It is a question that demands rules to answer it, rules that speak to that part of human behaviour which pertains to control over human behaviour itself, or at least some of the things that determine this behaviour. These rules, in turn, must be determined by some sort of values, values that say something about which things ought to shape our morals and how they ought to do this shaping.

Today, I think, we have more need of such rules than we used to. We have such need, I think, because in turning our ingenuity upon ourselves we have discovered – and continue to discover – fact after fact about how we may 'influence' human moral behaviour. Some such methods we have had access to for a long time; since antiquity we have been able to alter behaviour, more-or-less, through things such as force, fear and rhetoric[1]. Some such methods, though, we are only now discovering. The contemporary advertiser, in her most advanced form, augments the comparatively crude methods of the ancient orator with a corpus of techniques enlightened by a century of organised inquiry[2] and enabled by communication and data-gathering technologies barely a generation old[3]. The contemporary propagandist, sibling to the advertiser, renders the advertiser's new-and-improving tools into devices with which to make agents morally similar to himself or his paymasters[4]. Some contemporary medics, meanwhile, studying 'moral bioenhancement[5]' dream of methods of influencing which reach beyond the words, lights and sounds of time-honoured convention, and into the biochemistry of the brain itself, quietly altering the stuff of which agents – or at least, their embodied patterns of living – are made[6].

This is, then, a time of advancement in our methods of influencing human behaviour. These methods are getting better than they used to be. They are growing in the surety of their effects, the cheapness of their deployment and in their capacity to circumvent our attempts to evade them.

Often when influence happens unimportant things are influenced. The tools of the advertiser, for instance, are very often used to move us to make unimportant changes in which more-or-less identical products we consume. Influencing can also target, though, much that isn't unimportant. Some influencing effects, both potentially and indeed in fact[7], those special bits of ourselves we call

---

[1] G.A.Kennedy (1994), *A New History of Classical Rhetoric*, Princeton, Princeton University Press, pp.37-39

[2] A.Mackay (2005), *The Practice of Advertising [5th edition]*, Oxford, Elsevier Butterworth-Heinemann, pp.24-27

[3] L.A.Buchwitz (2018), 'A Model of Periodization of Radio and Internet Advertising History', *Journal of Historical Marketing Research*, vol.10, no.2, pp.130-150; pp.141-143

[4] *see for example* Y.Dai and L.Luqiu (2020), 'Camoflaged Propaganda: A Survey Experiment on Political Native Advertising, *Research and Politics*, vol.7, no.3, pp.1-10

[5] P.Crutchfield (2016), 'The Epistemology of Moral Bioenhancement', *Bioethics*, vol.30, no.6, pp389-396; pp389-390

[6] *Ibid.* p.390

[7] N.Levy, G.Kahane, P.Cohen, M.Hewstone and J.Savulescu (2014), 'Are You Morally Enhanced?: The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-125; pp.111-112, 122-123

our 'morals'. Some influencing changes the causes we support[8], what we are willing to do[9] and what we call 'good' or 'bad[10]'.

Though these 'moral dispositions' of ours (see §2.3 for a full definition) are sometimes protected from change by psychological or behavioural defences they exist within the causal systems that embed human minds and bodies. Morals can change; this we all know well. Morals can also *be* changed. Perhaps sometimes when morals change it seems spontaneous, a 'change of heart' that arrives without herald. More often, though, changes in morals seem to be caused things – in some sense – things where we can tell some plausibly true story about how the change in question came about. Such stories may trade in reasons and beliefs and be the sort of things that may be evaluated by the moral epistemologist[11], but they need not be this way. Sometimes such stories can only be told in terms of psychological and even physical events about which the epistemologist must be silent. Sometimes such stories must be made sense of in terms of the motley class of causes we call 'influences'.

One might think, for instance, of the infamous case of Aylan Kurdi, a three-year-old refugee who drowned crossing the Mediterranean in 2015. Pictures of his body circulated widely and caused an outcry which shifted the European debate about the plight of refugees significantly, weakening 'no rescue' advocacy at least for a time. Few, however, learned anything much from those pictures. The plight of refugees and their deaths and the deaths of their children crossing the Mediterranean were already well publicised. Something about seeing a dead child plucked from the sea, however, somehow motivated people more than their existing beliefs[12]. Something about that sight – something about those pictures – *influenced* those that beheld them, in a way which must be explained by reference to something other than altered beliefs or ideals.

One might think also, though, of the story of Phineas Gage. The workman – so the old neuroscientists' story goes – underwent a somewhat lasting and significant change in character after sustaining a severe brain injury[13]. He became altogether less easy to get on with in the aftermath of his accident, a development conventionally attributed to the direct effects of his injury[14]. In trying to describe this situation, we must say that Gage's injury had an *influence* on his behaviour – including his moral behaviour. Again, the important causal element in this story seems not to have been some change in Gage's picture of the world.

If morals may be *influenced*, then, and influenced through such non-cognitive causal means, this in-itself and apart from any special concerns of the present age warrants some sort of ethics. Morals are, after all, *important*. They, or at least the part of them I'll call moral dispositions, determine to

---

[8] J.R.DeCook (2018), 'Memes and Symbolic Violence: #proudboys and the use of Memes for the Propaganda and the Construction of Collective Identity', *Learning, Media and Technology*, vol.43, no.4, pp.485-504; p.490

[9] E.Ferrara (2017), 'Contagion Dynamics of Extremist Propaganda on Social Networks', *Information Sciences*, vol.418-419, no.1, pp.1-12; p.10

[10] W.J.Brady, M.J.Crocket and J.J.Van Bavel (2020), 'The MAD Model of Moral Contagion: The Role of Motivation, Attention and Design in the spread of Moralised Content Online', *Perspectives on Psychological Science*, vol.15, no.4, pp.978-1010; p.991

[11] A.Zimmerman (2010), *Moral Epistemology*, New York, Routledge; pp.1-2

[12] N.El-Enany (2016), 'Aylan Kurdi: The Human Refugee', *Law Critique*, vol.27, no.1, pp.13-15; p.15, L.G.E.Smith, C.McGarty and E.F.Thomas (2018), 'After Aylan Kurdi: How Tweeting About Death, Threat and Harm Predict Increased Expressions of Solidarity With Refugees Over Time', *Psychological Science*, vol.29, no.4, pp.623-634; p.631

[13] M.Macmillan and M.L.Lena (2010), 'Rehabilitating Phineas Gage', *Neuropsychological Rehabilitation*, vol.20, no.5, pp.641-658; p.642

[14] *ibid*. p.642

some extent how we live and how we feel about how we live[15]. They move us regularly to grand action or paralysis. Many of mankind's greatest achievements and worst blunders have been made in the name of this-or-that *moral* project and are only called great or awful viewed through some *moral* light. If such powerful and important things as morals – and moreover the moral dispositions that give them practical life – may be influenced, may be subjected to causal pressure from this-or-that quarter and change in response, then it's obvious that we should be careful about how we exert, manage and react to such influence. Giving meaning to precisely this sort of *carefulness* requires an ethics. It requires 'an ethics of moral-dispositional change'; a 'moral theory' describing the rights, wrongs, goods and bads of influencing peoples' moral dispositions.

Such an ethics by its nature must be a partial one. Though it deals with influences upon morals it cannot and should not pretend to comprehensiveness. Like an ethics of eating or accounting or leadership, such an ethics must limit its claims as far as possible to only concerns which arise specifically within the vicinity of its subject matter, and then perhaps not even all of these. There are, for example, concerns about the purposes to which influencing is put which may only be offered alongside a comprehensive ethics and hence cannot be usefully addressed as part of an ethics of influencing moral dispositions alone. Influencing someone to begin killing innocents or stealing pets, say, is a bad thing to do, but not because of any isolable ethics of influencing morals dispositions so much as ethics which make killing innocents and stealing pets bad things to cause or encourage. This is not to say that the rights, wrongs, goods and bads of influencing moral dispositions can or should be considered only in isolation, or that the concerns involved do not interact with concerns best considered in treatments of other parts of ethics (indeed, I think very often there is such interaction, and it is determinative of right action). It is to say, however, that there are at least some concerns specific to thinking about the ethics of influencing moral dispositions which may be profitably considered in isolation from concerns about what we are influenced towards or the consequences of influencing.

It is my view, in brief, that there exist certain things which – in a meaningful sense – have special rights to exert influence on morals dispositions. These things are such that their influence upon morals (without sufficient reason) ought not be opposed and ought to be protected from displacement by the influence of other things. We call these special things reasons, things which by their nature – different views of which I'm mostly neutral on – have content 'relevant' to the changes they inspire. When such content perceived through presentations or representations to an agent causes moral-dispositional change in said agent this change acquires a certain sort of merit. Wherever such content is absent or ineffective in an influencing of moral dispositions something goes wrong. This going wrong, in turn, gives us a pro-tanto reason to work against the influence in question. This may involve forgoing the use of said influence or resisting its use on others or oneself, provided there do not exist sufficient reasons to tolerate said use as a lesser evil.

For example, when should one give in to pressure to change one's ways and when should one resist this pressure? Suppose, for instance, I am an activist pro-lifer. I vociferously oppose expansion of access to abortion, tell anybody who will listen about my views, vote for candidates who share my position on the matter and refuse to facilitate abortion in my work life. Suppose, as it happens, that I have some close friends who abhor my activism, constantly complain about it, and begin to shun me on account of it. If I choose to maintain the friendships in question, over time, by the action of a subtle social pressure to coordinate one's behaviour with one's peers, I can reasonably expect my

[15] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; p.6

pro-life activism to diminish or even disappear[16]. Alternately I could stop maintaining the relevant friendships, or perhaps take steps to somehow quarantine my activist dispositions from these friendships so that the latter cannot influence the former. What is one to do in such a situation? Ultimately a full answer to this question depends partly on the evaluative qualities of my own behaviour; am I taking a valiant moral stand or am I fighting for a malevolent cause? How valuable my friendships are and what is possible within their context should probably also count for something in my judgement. What also must count for something in a case like this, though, is whether I am willing to tolerate the *way* in which I will be changed if I maintain the status quo of my friendships. Do I want to be moved to become less activist by the kind of 'peer pressure' and habituation which I foresee will change me if I proceed without curtailing or modifying my interactions with my friends?

One might think also of similar questions that catch us in relation to other influences we encounter in life. Many of us take special behavioural and psychological steps to limit or resist compunctions introduced (more or less successfully) by even quite innocuous advertising, and certainly 'propaganda[17]'. We may avoid and skip adverts when we get the chance, avert our attention or intentionally subvert their emotional content[18]. We do these things precisely because we are uneasy with the influence over our behaviour the advertiser and propagandist seek and sometimes procure[19]. Nonetheless such 'resistance' steps are – often not accidentally[20] – inconvenient or costly to execute, and we may wonder whether they are really justified. Answering such wondering requires, though, some sort of standard for evaluating influences, and deciding how much (if anything) ought to be sacrificed to render them ineffective. This, in turn, requires an understanding of precisely how influences may do harm, including how they may harm us in virtue of the methods of influencing they employ.

Questions which demand such evaluation of influences can also assail us when we assume the role of influencer ourselves.

In our lives, quite frequently, we influence others. Sometimes this is more-or-less accidental (as with, say, the social pressure exerted by the friends of my putative pro-lifer) but it may also be intentional. We may, for example, take steps to look a certain way so as to influence others towards becoming positively disposed towards us and our suggestions. We may tell our children the myth of Santa so they might associate certain behaviours with rewards in the future. We may employ tricks of rhetoric (such as listing things in threes…) to add 'gravitas' to the things we say. Sometimes, when we reflect on such influencing, we confront a certain unease. We wonder whether these sorts of influencings are quite *fair* to those on their receiving ends, whether in influencing people using such methods we might sometimes, somehow, harm them.

We wonder this most when we use such methods of influencing to try to change the special and important parts of others that shape their moral behaviour – their 'moral dispositions'. It is one thing to dress well in order to sell somebody shoes, say, it's quite another to dress well in order to get somebody to rob a bank. Intuitively, in the latter case, if how you look determines the way the subject of your influence behaves (as one might expect it to, at least a bit, in the right circumstances)

---

[16] *ibid*. p.177

[17] M.L.Fransen, E.G.Smit and W.J.Peeter (2015), 'Strategies and Motives for Resistance to Persuasion: an Integrative Framework', *Frontiers in Psychology*, vol.6, article 1201; p.1

[18] *ibid.* pp.2-3

[19] *Ibid*. p.1

[20] L.A.Buchwitz (2018), 'A Model of Periodization of Radio and Internet Advertising History', *Journal of Historical Marketing Research*, vol.10, no.2, pp.130-150; pp.142-145

then something rather more tragic will have happened. This relative tragedy cannot be just a matter of consequences for it remains even if, as it happens, the subject of your influence doesn't rob any banks (suppose he can't find any). Something about your dressing well leaving somebody disposed to rob banks seems bad in a way that your dressing well leaving somebody disposed to buy shoes is not, at least to the same extent.

This awfulness is not, furthermore, entirely a matter of the awfulness of the aims thus generated. There's something at least a little bit troubling about your dressing well causing somebody to become disposed to give to charity, say, at least relative to achieving the same results by – for instance – showing someone the good their charity could do (this is a species of example I will explore in detail in Chapter 3). This isn't to say that we can't dismiss such qualms for the sake of greater goods (to be got, for instance, through increased charity donations) but it is to say that we *do have such qualms*, and that they seem to vary by methods of influencing, not just effects (they occur whether-or-not influencing produces improvement in moral dispositions). This suggests there is a complexity to the ethics of influencing moral dispositions that requires understanding. Part of this understanding, furthermore, the part I will focus on in this thesis, can and must be neutral on the evaluative qualities of the moral dispositions we are influenced towards or away from; it asks not 'how can morals be improved?' but rather 'whatever your view is of moral improvement how ought it be achieved?'

We also – for we are creatures of moderate internal conflict – can find ourselves wishing to influence ourselves and confronting similar problems. Pascal said that those that wish for faith should act as if they have faith and thus become faithful[21]. Faith aside and humans being creatures of habit, something resembling this trick certainly works on a wide variety of behaviours. It is a commonplace that if you act a certain way enough eventually you'll become disposed to keep acting in that way in general. In deciding whether to employ this method in order to shape one's dispositions, though, one might still have second thoughts. One might wonder whether merely getting into a habit is the right sort of way to become disposed to do certain things, especially if these things have a moral quality. Consequences speak for themselves, but we might still feel like we want something more from our sources of moral-behavioural inspiration. Some criticisms of 'nudge theory', which champions a variety of subtle means of shaping 'choice architecture' so as to habituate good behaviour, trade on precisely this sort of intuition[22]. Similarly, one might think it is possible to have 'too much willpower', be so good at influencing one's own moral dispositions by focussing on one's beliefs that one renders oneself unresponsive to things that really ought to influence one's behaviour. This was how Jonathan Bennett analysed the holocaust planner Heinrich Himmler, anyway[23]. An agent with excessive willpower of this sort is a highly efficient influencer of him- or her- self, and this seems to be a kind of fault, one which leaves them somehow lesser[24].

The general point here is that questions about the acceptability and relative justifiability of influences on moral dispositions are ubiquitous within life. As I've suggested, though, in the present age they are beginning to be both forced and radicalised by advancing technology. For many of us, now, various media devices and platforms have more and better funded access to us than our

[21] M.Moriarty (2020), *Pascal: Reasoning and Belief*, Oxford, Oxford University Press; pp.368-369
[22] C.Sunstein (2015), 'The Ethics of Nudging', *Yale Journal on Regulation*, vol.32, no.2, pp.413-450; pp.427-428
[23] J.Bennett (1974), 'The Conscience of Huckleberry Finn', *Philosophy*, vol.49, no.188, pp.123-134; pp.127-129
[24] *ibid.* pp.133-134

friends, families and peers ever will[25]. Most of these systems are constructed and maintained by the enormous advertising businesses that populate the contemporary internet. These businesses – among other things – facilitate and sell influence over their users, building upon decades of research into how one might influence somebody towards purchasing shoes, saving for a car, or supporting a political movement[26]. They are influencers corporate and clever but are no less influencers for these facts. Their influencing is thus – at least when it targets dispositions with some moral quality – subject to all the same questions we may ask generally about the acceptability of such influencing, in a form radicalised by high stakes.

(1.2) Towards an Ethics of Moral-dispositional Change

Questions about the ethics of influencing have long found a home within the extant literature on the ethics of advertising[27], and also within literatures on means of social change which take inspiration from insights employed within advertising (think, for instance, of debate about the ethics of 'nudging[28]'). Such literatures have historically been dominated by debates between supporters of the value of autonomy, who take issue with various advertising methods owing to their use of sentiment[29], and consequentialists, who appraise the same methods in terms of their various results[30]. The former often commit to what Gibbard called a 'Protestant-like independence of mind[31]', opposing all sufficiently exogenous influences on behaviour, while the latter generally worry about the need to preserve social goods allegedly fostered by the advertising industry (such as economic growth)[32]. Comparable questions have more recently begun to populate a comparatively young literature on the topic of 'biomedical moral enhancement'. Within this literature biomedical techniques – which approach plausibility now more than in the past – are postulated which have the capacity to somehow change the morals agents act upon. This literature challenges us to judge whether we should use or even compulsorily deploy such techniques for the betterment of humanity. As with the extant literature on the ethics of advertising the resulting debate tends towards mainly consequentialists defending biomedical moral enhancement against other thinkers, often proponents of some particular parts of the ethics of moral-dispositional change, who criticise it in practice and principle[33].

What is common to these debates is that they seek to uncover what things ought to influence us and why – what methods of influencing are such that they ought (or ought not) to shape our characters and, sometimes, our moral characters.

---

[25] D.Aizenkot (2020), 'A Quantitative and Qualitative Approach to Analysing Cyberbullying in Classmates' Whatsapp Groups' in D.Nguyen, I.Dekker and S.Nguyen (2020) [eds.], *Understanding Media and Society in the Age of Digitalisation*, Cham, Palgrave Macmillan, pp.185-208; pp.188

[26] L.R.Samuel (2013), *Freud on Madison Avenue: Motivation Research and Subliminal Advertising in America*, Philadelphia, University of Pennsylvania Press; pp.183-188

[27] M.E.Drumwright and P.E.Murphy (2009), 'The Current State of Advertising Ethics: Industry and Academic Perspectives', *Journal of Advertising*, vol.38, no.1, pp.83-108; p.84

[28] C.Sunstein (2015), 'The Ethics of Nudging', *Yale Journal on Regulation*, vol.32, no.2, pp.413-450; pp.427-428

[29] A.Villarán (2015), 'Irrational Advertising and Moral Autonomy', *Journal of Business Ethics*, vol.144, no.3, pp.479-490; pp.483-485

[30] *ibid*. pp.482-483

[31] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; p.177

[32] A.Villarán (2015), 'Irrational Advertising and Moral Autonomy', *Journal of Business Ethics*, vol.144, no.3, pp.479-490; pp.482-483

[33] *see* J.Specker, F.Focquaert, K.Raus, S.Sterckx and M.Schermer (2014), 'The Ethical Desirability of Moral Bioenhancement: A Review of Reasons', *BMC Medical Ethics*, vol.15, no.1; pp.67, 10-11

We are awakened to the need to consider such matters by an awareness that there is a very real risk that at least some methods of achieving dispositional change transgress some unclear but important ethical line[34]. 'Brainwashing' and its cognates strike us as real and disturbing possibilities in our thinking about how morals may be changed, and though what 'brainwashing' is is not clear it clearly seems to have something to do with the methods used to influence those agents subjected to it. Drawing the line between such abuses and the realm of acceptable persuasion must be achieved before consequential and other concerns are factored into judgement. If a method of moral-dispositional change is proposed as a lesser of evils, we must first say something about whether it is and to what extent it is indeed an evil and why. Saying such things, never mind situating them within a wider ethics which contains other concerns, demands an understanding of what it is that can make a thing an acceptable influencer of morals. Attaining such understanding – in a general form fit to be applied to relevant thinking within varied literatures – is precisely the job of the 'ethics of moral-dispositional change' I'll try to articulate.

Within metaethics there have been some attempts to speak to such a comprehensive perspective on the ethics of influencing moral dispositions which I will draw upon and critique. Allan Gibbard, in facing the question of whose influence upon our behaviour we should accept, encountered the further question of what can make influences acceptable in general. His primary answer, citing 'indirect' pragmatic content and supported by an argument from scepticism[35], is one I'll criticise in §3.4. Brandt, meanwhile, answered the same sort of question by theorising that some influences hold a special 'relevance' to certain dispositions and that this relevance entitles them to exert influence upon these dispositions[36]. I am sympathetic to this understanding, as applied to the problem of discerning acceptable moral-dispositional influences rather than Brandt's original problem of defining what's good for an agent[37]. It is from exploring this understanding that my positive account of the ethics of moral-dispositional change will emerge.

Understandings of freedom, in turn, provide this positive account with a useful foil. These understandings, whether occurring within debates about the ethics of advertising[38], moral bioenhancement[39], or interpersonal power have always had something to say about the ethics of influencing. I will not base my position on claims about the nature of freedom itself, but I must address at least the most relevant of such claims as part of my attempt to describe the ethics of influencing moral dispositions. I'll respond specifically to two distinct understandings of freedom in offering my account: freedom as autonomy and freedom as interpersonal non-domination. The former, and its interaction with the ethics of moral-dispositional change, I'll understand through the works of Kant and his successors. The latter interpersonal sort of freedom and the claims it makes on our evaluation of influences I'll understand through ideas like those of Berlin[40] and Petit[41].

(1.3) Summary Positive View

---

[34] *ibid*. pp.14-15

[35] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; pp.179-181

[36] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313, R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.111-112

[37] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.353-354

[38] *see for example* A.Villarán (2015), 'Irrational Advertising and Moral Autonomy', *Journal of Business Ethics*, vol.144, no.3, pp.479-490; pp.483-485

[39] *see for example* T.Douglas (2014), 'Moral Bioenhancement, Freedom and Reasoning', Journal of Medical Ethics, vol.40, no.6, pp.359-360; p.360

[40] A.S.Kaufman (1962), 'Professor Berlin on 'Negative Freedom'', *Mind*, vol.71, no.2, pp.241-243; p.241

[41] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; p.55

In this thesis I argue that when we evaluate ways of influencing human moral dispositions one thing which matters is that these ways work by representing or presenting to influenced agents reasons which count in favour of the dispositions said agents are being influenced towards. Ways which work in this way thus deserve respect and protection and ways which don't are bad and harmful and deserve to be resisted, avoided and opposed. These things are the case, at least, provided there are no sufficient reasons to act otherwise, as in practice our responses to influencings should also be guided by instrumental considerations to do with the effects of dispositional changes (upon agents' practiced morals, more generally, and perhaps also their coherence[42]).

This standard applies to all cases of the influencing of moral dispositions, for better or worse, through accident or design, by man or nature. It persists however reasons are presented or represented to us such that they effect our dispositions (whether this be by beliefs, images, speeches, songs or whatever), so long as it is this representation or presentation that is doing the disposition-changing work. This standard allows for the influencing of dispositions through emotional arousal, provided this arousal appropriately connects reasons to dispositional changes (as it sometimes does). Wherever minds cannot facilitate such connections (as in young children, or the sufficiently irrational) my account cannot be applied.

I believe that by evaluating moral-dispositional influences according to their use – or lack thereof – of representations or presentations of such 'relevant' reasons we can protect a state that we value. We wish (all other things being equal) in at least our moral behaviour to be 'reason-responders[43]' – *to be moved to action only by the contents of reasons* – and we may only accomplish this insofar as we develop dispositions by perceiving and responding to the contents of reasons relevant to these dispositions.

I thus suggest an ethics of influencing moral dispositions which holds that working via the presentation or representation of relevant reasons is always a pro-tanto better-making feature of moral-dispositional influences (it gives their influence a certain kind of pro-tanto legitimacy). I call this a 'rationalizability requirement'. In my view influencing moral dispositions using reasons – the more the better – is always in at least one pro tanto sense good, influencing in any other way is always in this same sense bad.

I suggest, further, that the ethical significance of these influences and hence the force of the demands made by the rationalizability requirement are proportioned to the importance of dispositions being influenced, the volume of dispositions being influenced (which may be spread across multiple people) and the effectiveness of the influencing deployed. The more significant an influence is by this rough calculus the more important it is that it works by showing its subjects relevant reasons, the worse it is when it does not (ceteris paribus).

Such proportioning of the demands of the rationalizability requirement is significant because these demands are defeasible; they may be overridden in circumstances where there exist sufficient

[42] That is, whether agents act as they believe or perhaps feel they ought to, as specified by their 'moral ideals' (see §2.3).

[43] This is not the same thing as being reason responsive. To be a reason-responder is to be such that when you act you do so in appropriate response to reasons; to be reason responsive is to be such that you respond appropriately to the reasons you encounter. Plausibly, you must be the latter to also be the former, but you need not be the former if you are the latter. You might be reason responsive but not a reason-responder inasmuch as when you act you do so in response to non-reasons (you might, say, have been influenced to do something by the administration of a drug that generates a compulsion to do the thing without in wany way showing you a reason to do it), even though you would respond properly were you to encounter any reasons.

reasons to, say, make people better through means of influencing other than showing them reasons, or to stop them being made worse by perception of reasons to become worse. Inasmuch as this is the case my account generates only pro tanto reasons to alter the ways in which we influence or tolerate influencing. Nonetheless these reasons are not insignificant; where all other things are equal they should be determinative of right action, as they may also be whenever other considerations are not strong (or, in practice, clear) enough to be decisive or the demands of the requirement have particular force (as they do when many are influenced to make major changes in their moral dispositions, say).

I further contend that whether one valorises freedom as autonomy (construed as requiring action structured by a 'moral incentive') or as interpersonal non-domination one should find at least practical agreement with the claims of my account. If one cares about these things, I argue, one ought to apply the standard I supply to one's evaluations of influences on moral dispositions. This being the case this standard helps unify, at least in practice, some of the approaches to evaluating moral-dispositional change advocated in extant literature.

It is my belief that the moral theory of influencing moral dispositions – the 'ethics of moral-dispositional change' – that I offer articulates a neglected part of our thinking about the rights, wrongs, goods and bads of influencing morals. This result, in turn, ought to help inform all those whose business is such influencing or the ethics and regulation of such influencing, or indeed life in a society where such influencing can occur.

(1.4) Thesis Summary

I'll begin my substantive project by defining its terms and scope in Chapter 2. This chapter characterises the 'moral dispositions' I'm concerned with. It shows how they may be distinguished from non-moral dispositions which govern behaviour of systematically less important sorts and 'moral ideals' which don't determine moral behaviour in the same direct way. Chapter 2 argues that, in assessing the ethics of influencing moral dispositions, it's necessary to distinguish between considerations distinctive to this ethics and 'instrumental' considerations of a more general sort, and it shows how this may be done. It justifies the general focus on the former sort of considerations in this thesis, arguing that any 'instrumental' analysis of the ethics of influencing moral dispositions exclusively instrumental enough to ignore distinctive considerations may only be evidenced by resisting disconfirmation by inquiries such as the one I undertake. While developing these foundational distinctions, Chapter 2 lays further groundwork for my argument by defining important recurring terms such as 'effectiveness' as applied to methods of moral-dispositional change.

Chapter 3 introduces a thought experiment to tease out intuitions upon which to build my ethical model. In this thought experiment two maximally similar characters are posited who differ only in their methods of moral-dispositional change, whereby one influences their own moral dispositions by exposing themselves to representations of certain features of reality while the other accomplishes the same influential task by exposing themselves to inspiring but vacuous rhetoric. We favour the former method of influencing, I suggest, but it isn't clear why. Potential explanations of this thought experiment based on pragmatism[44], transparency[45] and informed non-resistance[46] are

[44] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; pp.223-226

[45] *see* C.Cowley (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290

[46] *see* J.Christman (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy*, vol.21, no.1, pp.1-24

explored and rejected in favour of an understanding which valorises connections between human moral behaviour and certain 'relevant' parts of reality. I then attempt to give meaning to this 'relevance' and thereby offer the theoretical core of my account of the ethics of moral-dispositional change. Analogising from Brandt's cognitive psychotherapy[47] and the concept of 'logical relevance' he cited in debating Stevenson[48], I argue that 'relevance' in the relevant sense must be interpreted as a form of rationalizability constraint. This interpretation informs my core positive claim, that legitimate (in a certain pro tanto sense) influencing of moral dispositions requires that they only be influenced by presented or represented reasons. I call this the rationalizability requirement. This view, I show, resists criticisms levelled at Brandt's analogous position, notably by Velleman[49]. Concerns about the underdetermination of dispositional change by experience[50] are undercut by the incorporation of 'vividness' as a modifier on the legitimacy of influences (an incorporation I may accomplish, working posterior to accounts of reasons themselves, in a way that Brandt could not) and – contra Velleman[51] – facts can associate with particular motivational impacts (at least in such a way as to overcome his 'problem of representation').

Chapter 4 seeks to move from the theoretical account of the ethics of influencing moral dispositions articulated in Chapter 3 to a maximally useful understanding for practical ethics. In the first half of Chapter 4 I show how, despite disagreement and uncertainty about reasons, the existence of obvious reasons and non-reasons as well as non-representational or (sufficiently) poorly representational influences creates room for agreement about the evaluation of influences given the rationalizability requirement. This is the case even if weak reasons are ubiquitous and we do not always respond proportionately to reasons. In the second half of Chapter 4 I proceed to show, at least in a preliminary way, how the demands of the ethics I propose may be weighed against other ethical demands we encounter in life and which can conflict with the demands of ethics distinctive to changing moral dispositions.

In Chapter 5 I clarify and toughen my position by comparing it with extant positions which stress the value of freedom, understood either as autonomy or as interpersonal non-domination. Targeting Kant's autonomism as a useful reference point for the wider tradition, I argue that my position is compatible with an interpretation of autonomy which identifies ideal agency with responding to the reason-making contents of practical reasons. I argue that alternative interpretations of the autonomy tradition incompatible with my positive account are vulnerable to charges of 'moral fetishism' of the sort popularised by Michael Smith[52]. In doing so I defend a particular view of ideal agenthood. This ideal agenthood, in turn, needn't be understood in terms of beliefs and their relations with behaviour (contra Kant, on some readings[53]); any successful representation or presentation of a relevant reason may legitimately influence moral dispositions. The best autonomist accounts of the ethics of moral-dispositional change, thus, ought to embrace my rationalizability requirement.

Recognising that the concept of autonomy can't exhaust the language of freedom as it applies to debates about the acceptability of influences, I also discuss interpersonal freedom. Arguing that one

---

[47] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.11

[48] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313

[49] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.357-371

[50] *ibid.* pp.361-362, 365-368

[51] *ibid.* p.366

[52] M.Smith (1994), *The Moral Problem*, Malden, Blackwell; pp.71-76

cannot dominate another by showing them reasons (even in cases where things might seem to be otherwise), I suggest that my rationalizability requirement covers all ground that ought to be covered by applying a theory of interpersonal freedom (such as the ones articulated by Berlin[54] or Petit[55]) to the evaluation of moral-dispositional influencing. Hence, any value accorded to interpersonal freedom supplies reason to apply the standards I articulate to the evaluation of moral-dispositional influencing.

Given all of this, if you value the preservation of freedom (as many seem to) through processes of influencing – in some of its most popular senses – you ought to embrace the rationalizability requirement.

The bridge into practical ethics prepared in Chapter 4 and the peace brokered with extant literature in Chapter 5 together create space to apply my account in critique of extant and future practices. This space is explored, in a preliminary way, in Chapter 6. In this chapter I argue that my account has important implications for how we manage the influencing of moral dispositions. It shows, I think, that means of influencing that fail to present or represent reasons ought to be opposed, something which raises the burden of moral proof required to use such techniques as part of campaigns of biomedical moral modification, 'nudging' or propaganda. I propose, indeed, that at least some techniques in use in such fields (most significantly in advertising) have a systematically unethical character, given the rationalizability requirement, and must be sufficiently justified or else opposed. My account also shows that certain influences ought to be accorded a certain respect and should not be purged from development of human moral dispositions save when this is needed to prevent some greater evil of people being led astray. This all being the case, my account demands a complex mix of ignoring, attention paying, reconsidering, and perhaps legislating. If we satisfy these demands, though, the valuable thing that is living one's morals in response to one's reasons should be able to persist perhaps even into a future of ubiquitous manipulation of moral dispositions.

---

[53] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; pp.30-34

[54] I.Berlin (2002 [1957]), 'Two Concepts of Liberty' in H.Hardy [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217; pp.169-170

[55] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; p.55

Chapter 2: Foundations

(2.1) Abstract

This is a chapter of foundations. It defines the scope and terms of my thesis, differentiating the special class of moral-dispositional changes in agents that I will be concentrating on from non-moral and non-dispositional changes. In doing so I separate my project from moral epistemology. I narrow my focus onto concerns regarding the means by which moral-dispositional changes may be caused. Competing instrumentalist accounts of the ethics of changing moral dispositions may be either moderate or extreme. Moderate instrumentalist accounts, which hold that the results (however one evaluates them) of moral-dispositional changes matter, do not compete with but rather supplement understandings of how moral dispositions ought to be changed. Extreme instrumentalist accounts, on the other hand, may only be evidenced by the absence of non-result factors in the proper evaluation of moral-dispositional influences. Their truth is thus best determined by potentially disconfirming investigations such as this project. Methodological commitment to reflective equilibrium and the distinction between inter- and intra- personal concerns about the ethics of moral-dispositional change are highlighted.

(2.2) Moral Change

If I am to say anything about the ethics that govern our judgements about moral-dispositional changes, I must say what moral-dispositional changes themselves are. Over time I may go from favouring blue to favouring yellow, I may go from being 183cm tall to being 180cm tall, or I may go from giving a large proportion of my income to charity to relatively little. These are all changes. The lattermost change, I suggest, has a moral character. The other changes don't. What makes the lattermost change in this way different to the other two? What is *distinctive* about changes like the lattermost change that makes them 'moral' in character?

This is a particular instance of the broad problem of identifying what distinguishes moral things from non-moral things. What makes moral evaluations, say, different from the evaluations involved in etiquette, and so on[56]? Without offering some sort of answer to this broad problem, no clear 'ethics of moral-dispositional change' may be stated. Without meaning to wade too deeply into this broader problem, I suggest that the distinctive feature of moral things is their possession of or special relation to a certain sort of normative authority. I believe that this understanding of what it is to be 'moral' is best suited to ground my project and focus it on the phenomena I wish to interrogate, at least.

To flesh this out in terms of reasons, as a start, a non-moral reason may or may not be outweighed by any other non-moral reason, but a moral reason outweighs any non-moral one and may only be outweighed itself by another moral reason in the determination of right action. 'Right' here is understood in *sui generis* terms; what you ought to do in the given case, given everything.

It may be good etiquette to slurp drinks around here, but I may be uncomfortable slurping my drink and thus have a prudential reason to avoid doing so. The balance of right judgement may sit with either course of action in a case like this, but it isn't resolved at the outset by the nature of the reasons involved. It rather must be discerned based on how important the reasons in play are

[56] D.Dorsey (2016), 'Moral Distinctiveness and Moral Inquiry', *Ethics*, vol.126, no.3, pp.747-773; p.747

relative to each other. Which has the greater weight of reasons behind it; slurping or not-slurping? Any moral reason in such non-moral considerations, however, by its nature trumps the other reasons in play. Suppose it turns out – somehow – to be morally wrong to slurp one's drinks. I suggest that if this turned out to be the case then it would always be right to not slurp one's drinks no matter the balance of purely non-moral reasons. If I visited Russia, where slurping is the polite thing to do, and attended a state dinner where the etiquette stakes are as high as etiquette stakes get, I'd still be doing something wrong if I slurped my drink.

What's right is only determined by non-moral reasons when there are no moral reasons involved. When there are moral reasons involved the balance of these reasons alone determines what's right; non-moral reasons aren't outweighed in such considerations so much as side-lined, as if they don't involve a commensurable species of value. Perhaps they don't, or perhaps they do yet somehow moral reasons systematically carry a higher magnitude of weight. This capacity to side-line, 'override' – however it works – is what the special normative character of moral things consists in.

Dale Dorsey argues against this kind of story[57]. He argues that one can't distinguish the moral using normative authority in this way because the proposed normative authority of morality requires further explanation, and thus the way of distinguishing the moral that I propose must reduce to some alternative method[58]. Dorsey thinks this because he rejects the idea that it can just be a conceptual truth about the moral that it has special normative authority in the way that I describe. The special normative authority can't be merely part of what it is to be 'moral[59]'.

Dorsey, however, muddles the project of defining the moral in general with the separate project of applying this definition. If it were indeed a conceptual truth that what distinguishes the moral from the non-moral is a special relation to a kind of normative authority (as I propose it is), we might still ask for what things this relation obtains. If moral reasons have 'overriding' normative authority of the kind I describe, we might still ask 'what reasons are the moral ones?' This is a question about applying a general concept rather than a question about the concept itself. What we must determine is what specific properties of things fit them to attain the special moral sort of normative authority. This is the business of first order ethics, in which properties like 'will maximise happiness' or 'keeps a promise' are suggested as sources of the overriding moral sort of normative authority. They are suggested as sources of this sort of authority because it is understood a priori that were they to be sources of this sort of normative authority then they would be sources of distinctively moral reasons.

Think of mechanical hardness. It's a conceptual truth about 'hardness', as used in engineering (and less precisely in other settings), that hard things have a certain macrophysical character. They resist indentation, plastic deformation, and so on[60]. Knowing this, we might still ask 'what things are hard?' or 'what things are hard to some given extent?' (hardness being quantifiable). In answering this question we discover that there are many ways of being hard. Steel is hard by being composed of interlocking grains of metallically bonded iron. Diamond is hard by being composed of single large molecules of covalently bonded carbon. Corundum is hard by being composed of a lattice of aluminium and oxygen atoms ionically bonded into molecules which are electrostatically attracted to each other. The existence of such distinct ways of being hard creates no pressure to reduce the

---

[57] *ibid*. pp.765-767
[58] *ibid*. p.766
[59] *ibid.* p.767
[60] W.W.Geberich, R.Ballarini, E.D.Hintsala, M.Mishra, J.Molinari and I.Szlufarska (2015), 'Toward Demystifying the Mohs Hardness Scale', *Journal of the American Ceramic Society*, vol.98, no.9, pp.2681-2688; p.2681

category of 'the hard' to a story about these distinct physical properties. Rather, these physical properties are understood as particular ways of satisfying the conditions needed to be 'hard' to this-or-that extent.

Analogously I suggest that the properties first order ethics identifies are proposed ways of satisfying the conditions (attaining the special class of normative authority) needed to be 'moral'. We can thus generally distinguish the moral from the non-moral, just as we can generally distinguish the hard from the not-hard, whatever we say about particular cases. Note that this view needn't commit to any claims about how – or which – properties may come to grant this special 'overriding' normative authority. It just proposes that it is this overriding character that distinguishes the moral. Nor does this account suggest we can't go wrong in attributing this character.

How, then, can this broad understanding of distinguishing the moral be applied to the events in human lives called 'changes' that I'm interested in?

As suggested in the previous chapter, I'm interested in when people change their practised morals. I'm interested in 'influences' that go between acting as if it's right to X and acting as if it is not right to X (or the other way about). As such my interest is ultimately in things that determine outward behaviour rather than mentation. Were an ingenious doctor to devise a device that effectively modifies your psychology and thus grants him command over your behaviour, I'd be interested in the doctor's device. This interest would persist however the device works; whatever it does with your thoughts, desires, habits and so on. So long as it changes your behaviour in some stable way and does so by altering whatever psychology in fact determines your patterns of behaviour.

Some behaviours, I suggest, have a moral character. We may disagree about which ones do, but some plausibly do. Intentional arbitrary killing of people, for instance, pretty universally gets moral wrongness attributed to it. We assign behaviours moral characters in this sort of way when they are such that the distinctly moral sort of normative authority usually attaches to reasons for or against them. Arbitrarily killing people generally has a moral character because there always seem to be moral reasons not to do it. Sitting down generally has no moral character because only very occasionally are there moral reasons for or against it. Giving to charity generally has a moral character because there are usually moral reasons for or against it.

The changes I'm interested in are the ones in stable moral behaviours and inherit their moral character from these behaviours. A change in the amount one gives to charity is thus a 'moral' change insofar as giving to charity is a moral behaviour and giving to charity is a moral behaviour insofar as it is a behaviour which there are usually moral reasons for or against. These moral reasons are 'moral', in turn, on account of their special moral sort of normative authority. This indirect multi-part relation connects my subject matter with the overriding normative authority that I suggest delineates the moral. To be a moral change, for my purposes, is to be a change in a behaviour of a sort usually informed by reasons with the moral sort of normative authority. It is thus possible for a 'moral change' to be for the worse or the better; becoming a robber is 'moral change', just not a good one. It's my hope that this definition will be intuitive for most readers. For the unsatisfied, though, it's possible to regard it as stipulation and read on. The processes – 'influences' – that go between acting as if it is right to X and acting as if it's not right to X in paradigmatically moral contexts are worth evaluating, whatever we call them.

(2.3) Moral Dispositions and Moral Ideals

One peculiar aspect of the story I intend to tell is its somewhat behaviourist character. I'm interested in evaluating the sort of causal processes that make one, say, enact more murder, less larceny or

more charity. One might think that this is an odd sort of thing to look at. One might worry that such a discussion will neglect considerations of how one should determine whether one ought to become more murderous, less larcenous or more charitable. One might think that there are *moral-epistemic* approaches to the evaluative work I mean to undertake. I don't think this is the case (at least unless moral epistemology is understood as involving more than the study of how we might find out what normative properties things have) and I will here show why.

We humans sometimes have a certain problem. It's a familiar, common problem, so I'll call it 'the common problem'. We can want to live according to a certain moral norm, or even moral system, yet fail to live as we believe we should. We might feel we ought to play fair, yet cheat. We might think we ought to defend ourselves yet in practice, cower. Generally, we can encounter situations where (for better or worse) our views about how we should act and live differ from our behaviour. We can fail to live out our ideals.

When we encounter this sort of problem, it's clear that we're encountering a sort of clash. We think we ought to behave in some way, but something is interfering with us changing our practices. The precise description of this clash is a matter for moral psychology. A Humean, after Frankfurt, may describe such a clash as being fundamentally between second and first order desires[61]. A Scanlonian may suggest an inability to 'see' reasons properly[62]. A Gibbardian could anatomise the conflict in terms of differing accepted and internalised norms[63]. However you describe such situations clearly enough there's such a thing as a way we think we should be, want to be, that's distinct from at least part of the psychology that constitutes our actual behavioural dispositions.

Let's, ecumenically, say that the distinction in question is between 'moral ideals' and 'moral dispositions'.

'Moral ideals' represent a set of supposed (by the agent who has them) morally better dispositions to do given things in given circumstances and having a 'moral ideal' necessarily involves having some motivation to become as the ideal says you should be. I take no position on the anatomy of moral ideals. They represent a version of yourself and they motivate you to make yourself like this version; I claim nothing about how they do this. They may be single mental states or comprise multiple states in loose confederation. One example of a moral ideal is the motivating desire to become a person who refuses to steal under any circumstances. Another example of a moral ideal is the motivating belief that one should actively support – or for that matter oppose – access to abortion.

'Moral dispositions' are actual dispositions to act of the sort that agents actually have, constituted by whatever features of agents' psychology – habits, desires, motivating concepts, etc. – actually constitute such things. One example of a moral disposition is habitually not stealing. Another example of a moral disposition is desiring so strongly that no abortions happen that one tends to protest abortion clinics. Another example might be judging that people should have access to abortion and being reliably motivated by this settled judgement such that one tends to protest those who protest abortion clinics. Moral dispositions may be distinguished from non-moral dispositions as per their indirect but specific relationship with the distinctive normative authority of morality as per §2.2 (and as such might be, in whatever ultimate sense you prefer, good or bad).

---

[61] H.Frankfurt (1971), 'Freedom of the Will and the Concept of a Person', *The Journal of Philosophy*, vol.68, no.1, pp.5-20; pp.12-14

[62] T.Shapiro (2009), 'The Nature of Inclination', *Ethics*, vol.119, no.2, pp.229-256; p.242

[63] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; p.71

My evaluative project will focus on processes of change in moral dispositions rather than processes of change in moral ideals.

Note here: there's a difference between being disposed to X and being motivated to X. The former constrains patterns of behaviour, the latter constrains one's psychology (which may sometimes itself constrain behaviour). Due to this one can't declare moral ideals a subset of moral dispositions. One may be motivated to change one's ways, and thus have a certain moral ideal, without also being disposed to change one's ways (perhaps because the motivation in question isn't strong enough to effect behaviour).

Generally, when we speak of peoples' morals we more-or-less equivocate between speaking of persons' moral ideals and their moral dispositions. Sometimes we mean to say things about the kind of people we believe them to aspire to be and sometimes (perhaps more often) we mean to say things about how they're disposed to act. Sometimes, plausibly, we also mean to say things about both.

Some don't distinguish between moral ideals and moral dispositions. Socrates, in *Protagoras*, denies that we can be motivated to do other than that which we find good. He suggested that to act as if one thinks something is good shows that one thinks something is good and *that's all there is to it*[64]. This simple idea enjoys intermittent popularity. Users of 'revealed preference theory' could be interpreted as embracing a Socratic position[65] (at least for empirical convenience[66]), and McDowell has offered a somewhat similar view[67]. If this position is right, then there's no useful distinction between moral ideals and moral dispositions. There's no such distinction because to have a moral ideal, to recognise a behaviour as superior, is – given this view – precisely to adopt the dispositions involved in said supposed superior way of living.

I've nothing to add to debate about the Socratic position besides agreement with the common-sense view holding that human moral psychology is more complicated[68]. For one thing, 'the common problem' is a problem that we experience in life and a problem that a credible moral psychology must allow. Denying the common problem seems to be denying real phenomena. This problem with the Socratic view – that it seems to deny obvious features of human moral life – is one that has attracted frequent notice. Authors since Aristotle have complained that the Socratic view fails to adequately account for 'akrasia', where one fails to do as one judges one should due to frustrating features of one's psychology[69]. I take it that this flaw in the Socratic position is lethal and that the species of akrasia (or 'akrasia of aim' in Rorty's terminology[70]) I've called the 'common problem' is a real problem we encounter.

It should be noted that the notion of 'moral dispositions' needn't be read as implying a commitment to a virtue-theoretic account of generalisable character traits. In any given individual moral dispositions *may be* like this, *or* they may be highly specific such that, say, a person lies to insurers but never to lawyers and cannot helpfully be called honest or dishonest in any entirely general

---

[64] J.Kennett (2001), *Agency and Responsibility*, New York, Oxford University Press; pp.9-12
[65] D.W.Hands (2013), 'Foundations of Contemporary Revealed Preference Theory', *Erkenntnis*, vol.78, no.5, pp.1081-1108; pp1087-1088
[66] *ibid.* p.1083
[67] J.Kennett (2001), *Agency and Responsibility*, New York, Oxford University Press; pp.17-25
[68] *ibid.* p.13
[69] *ibid.* p.14-15
[70] A.O.Rorty (1980), 'Where does the Akratic Break take Place?', *Australasian Journal of Philosophy*, vol.58, no.4; pp.333-346; p.333

sense. It's an empirical matter whether, for any given individual, she conforms to either one of these models of having moral dispositions and, indeed, for humans in general, which of these models is more common or perhaps universal. Inasmuch as this is the case the 'moral dispositions' I discuss should be compatible with the granular context-dependence of human moral behaviour proposed by situationalists[71]. Indeed I don't think one can really avoid having moral dispositions in my sense. Even one who behaves *randomly* in moral contexts still has the unusual moral disposition to 'behave randomly in moral contexts', a disposition which might change and thus generate space for the application of my evaluative project to said change. This point is further emphasized once it is remembered that moral dispositions as I discuss them need ground only tendencies and not inviolable action-determining rules. It is very hard to conceive of an agent who is not (among many other things) such that they *tend* to behave in some given set of ways in some given set of moral contexts.

Moral dispositions also shouldn't be understood as features of one's constitution so alien as to render actions they cause beyond one's responsibility. When one acts because of a moral disposition one has *one still acts* for the moral disposition is an integral part of the determination of one's behaviour, though it might itself be exogenously caused. This means that there may be 'mind control' cases (consider Persson and Savulescu's 'God Machine', which intervenes directly from the outside to alter specific intentions and behaviour[72]) *exclusive* of controlled persons' responsibility for actions undertaken under such control which cannot be understood as involving changes in moral dispositions. How one identifies such cases depends on claims about responsibility and its loss which I will not contest, though I do suggest, at least, that we may be responsible for the outworkings of our cognitions, desires, habits and heuristics. At least I shall select examples as if this is so. The reader may change out examples of moral dispositions as necessary should they believe any of these things incapable of causing responsible action in those possessing them. Moral-dispositional manipulation – what I discuss – preserves its subjects' responsibility for their actions. It works (when it works) by changing the agent who is manipulated, such that said agent still *does* what they do.

This being the case many standard 'precommitment' cases[73] fall outside my remit, for they don't involve moral-dispositional changes so much as responsibility shifts from future to present selves. Consider Ulysses lashing himself to his mast; this made his staying on his ship a choice of him before he heard the sirens' song and not him after hearing it. The only precommitment cases I may discuss are those in which precommitment works through some changing of moral dispositions (that is, in Elster's terms[74], *some* type '2' and '3a' cases). This being the case my claims about such cases shouldn't be taken to say anything about precommitment generally.

Given that moral ideals and moral dispositions can be distinguished, then, we must ask evaluative questions about processes of change in them separately. One can ask 'how should I go about changing my morals?', but this invites the question 'which bit of them; your ideals or your dispositions?' This is not a trivial question; it may be that changes in moral ideals and changes in moral dispositions should be evaluated differently in at least some respects. As our ordinary method of talking about 'having' morals equivocates between ideal and dispositional interpretations, it

---

[71] J.Sabini and M.Silver (2006), 'Lack of Character? Situationalism Critiqued', *Ethics*, vol.115, no.3, pp.535-562; p.537

[72] J.Savulescu and I.Persson (2012), 'Moral Enhancement, freedom and the God Machine', *The Monist*, vol.95, no.3, pp.399-421; pp.412-414

[73] J.Elster (1977), 'Ulysses and the Sirens: A Theory of Imperfect Rationality', *Social Science Information*, vol.16, no.5, pp.469-526; pp.471-475

[74] *ibid.* p.512

blinds us to this problem. By recognising the problem, though, we become able to give moral dispositions and moral ideals the specific evaluative treatments they require. This being said, it should be emphasised that both the ideals and the dispositions are part of what we normally call 'morals'. A career thief claiming ideals prohibiting larceny would be called hypocritical; someone habitually unable to steal confessing the belief that stealing isn't wrong seems somehow astray.

An evaluative project investigating the rights and wrongs of changing moral ideals is already well underway. This is, at least as far as ultimate objectives go, the essential project of moral epistemology. Moral epistemologies purport to reveal how we should discover what we should do or how we should live[75]. Insofar as they do this they propose norms by which we should shape change in our moral ideals; change in the things that codify for us what we should do and how we should live.

Moral-epistemic norms won't tell you, though, what to do once you've chosen a new way of acting or living; they only tell you how you should make the choice, how to properly select which ideals to hold. In principle a moral epistemologist – say one concerned with virtue – *could* advise you on what to do about these choices, once they're properly made, but insofar as they start doing this they'll be doing *more than just moral epistemology*. Pascal's advice, for instance, to behave faithfully so that one might find faith[76], isn't justified by the rather curious epistemic norms behind his famous wager – it's a separable part of Pascal's story justified by an isolable want for faith[77]. It's not at all clear how any moral-epistemic norm could be used to assess advice analogous to Pascal's to practice something (even something moral) so as to eventually become disposed to do it. Generally, norms governing how we should change how we act in moral contexts seem distinct from norms governing how we ought to decide how we should act.

(2.4) Evaluative Foci

If I'm to investigate the rights and wrongs and goods and bads of changing moral dispositions, then, I must first establish which features of such changes arouse concern.

Firstly, and most obviously, we can evaluate moral-dispositional changes in accord with their *content*. When someone goes from having the moral dispositions of an upstanding citizen to having the moral dispositions of a hitman, we're mostly concerned about the presence of another hitman in the world – because of the harm he might do or because of the badness of the moral dispositions thus created – and we evaluate involved moral-dispositional changes accordingly. This is the 'instrumental' part of the ethics of change in moral dispositions, the bit that's interested in what outcomes such changes deliver and evaluates these changes according to these outcomes. It asks whether moral dispositional changes make people better or worse, or result in more or less harm, or make people agree with us, or cost too much, or make peoples' dispositions cohere with their ideals, all depending on some background view of what is valuable (which may or may not cohere with some general moral theory like preference utilitarianism and may or may not be, inasmuch as such a thing can be, true).

'Instrumentalism' about the ethics of influencing moral dispositions, in turn, may be divided into two broad categories. 'Extreme Instrumentalism' holds that such outcome-evaluating can fully account for all norms that govern moral-dispositional changes. We want such changes to maximise value by

---

[75] A.M.Jaggar and T.W.Tobin (2013), 'Situating Moral Justification: Rethinking the Mission of Moral Epistemology', *Metaphilosophy*, vol.44, no.4, pp.383-408; p.385

[76] S.T.Davis (1991), 'Pascal on Self-caused Belief', *Religious Studies*, vol.27, no.1, pp.27-37; p.28

[77] *ibid.* p.28

cheaply producing better, less harmful, more agreeable or more coherent people and that's all there is to it. 'Moderate instrumentalism' on the other hand, an umbrella which covers most sensible views of the ethics of moral-dispositional change (including my own positive view, and plausible interpretations of at least the competing views discussed in §3.4, §3.5 and §3.6, perhaps even some interpretations of those discussed in Chapter 5), holds that outcomes matter to evaluating moral-dispositional influences, alongside other things. Understanding the character of these 'other things', though, nonetheless presents a specific evaluative challenge, one which will be my central task in this thesis.

Secondly, some authors attempt to evaluate changes in moral dispositions by examining the methods of change that are employed in them. They ask whether these methods do things like ignore important realities[78], track evolutionary imperatives[79] or compromise important freedoms[80].

Together such 'content' and 'method' theorists and their respective considerations capture most extant debate about the acceptability of moral-dispositional changes. There can be overlap in the recommendations of content and method theorists, with all judging that certain processes of moral-dispositional change are bad or wrong but differing in the reasoning behind these judgements.

There might also be other kinds of view on the ethics of changing moral dispositions. For example, one might insist that some rates of changing moral dispositions – not too fast or perhaps not too slow – are good or better than others in a way that's irreducible to claims about the contents or methods of dispositional changes. Dogmatists, who think that one owes loyalty to one's existing ways, seem to think something like this and there have been efforts to offer careful defences of this position with respect to moral conduct[81]. Valerie Tiberius, in justifying open-mindedness, has also offered an argument for a position in this vicinity, proposing that we ought to develop morals which are 'resilient' and thus to some extent stable to reduce the risk moral-dispositional change poses to our long-term projects[82] (although Tiberius' overall position is content-orientated, inasmuch as it places substantive constraints on the dispositions we ought to maintain[83]). I'll not address such 'rate' accounts at length. If you believe that there exists some better or right rate of moral-dispositional change, then you should be able to insist that change takes place at this rate irrespective of the claims about the ethical acceptability of this change offered in this thesis.

This project will focus on concerns about methods of moral-dispositional change. It asks, in general, 'in what ways should our moral dispositions be changed?' In doing so it means to ask something about which sorts of influences on our moral dispositions we should (ceteris paribus) tolerate and encourage and which we should (ceteris paribus) limit and resist.

I have two reasons for adopting this focus. Firstly, we are currently confronted with a glut of new methods of changing moral dispositions, as I stressed in §1.1. If they have any distinctive evaluative qualities, we need some way to spot them. Secondly, I suggest that content theories of the ethics of

---

[78] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.156
[79] A.Gibbard (1990[1992]), *Wise Choices, Apt Feelings*, Cambridge MA, Harvard University Press; pp.223-224
[80] J.Harris (2016), *How to be Good: The Possibility of Moral Enhancement*, Oxford, Oxford University Press; pp.80-84
[81] *see for examples* R.M.Adams (1995), 'Moral Faith', *The Journal of Philosophy*, vol.92, no.2, pp.75-95; pp.86-87, M.Pianalto (2011), 'Moral Conviction', *Journal of Applied Philosophy*, vol.28, no.4, pp.381-395; pp.381-389
[82] V.Tiberius (2012), 'Open-Mindedness and Normative Contingency' in R.Shafer-Landau (2012), *Oxford Studies in Metaethics Volume 7*, Oxford, Oxford University Press; pp.191-192
[83] *see for example* V.Tiberius (2018), *Well-being as Value Fulfilment: How we can Help Each Other to Live Well*, Oxford, Oxford University Press; pp.74-76

moral-dispositional change – at least in extreme instrumentalist form – may only be evidenced by the exploration of alternatives. 'Moderate instrumentalist' content theories, on the other hand, are consistent with accounts like the 'method' account I'll propose.

As I've noted the effect of integrating content concerns, about whether morals get better or worse (by whatever lights we illuminate such matters), into thinking about the ethics of influencing moral dispositions is to introduce some sort of instrumentalism. According to this *instrumentalism*, whatever form it takes and whatever values underpin it, moral-dispositional influences may be sensibly evaluated in terms of their results. These, in turn, may be understood virtue-theoretically in terms of the value (or disvalue) of the moral dispositions that agents acquire by way of whatever influences are at issue (Are they better moral dispositions? Are they moral dispositions consistent with ideals licensed by a sound moral epistemology? Are they moral dispositions coherent with effected agents' ideals?), or consequentially in terms of practical results, or indeed pragmatically in terms of – say – whether influenced morals cohere with influencers' agendas (if such considerations can have ethical force). Either way no evaluative claims are made about moral-dispositional changes in virtue of how they work; claims are only made in virtue of what they achieve.

Such analysis has a proper place in just about every evaluation of a moral-dispositional influencing. Clearly when we evaluate moral-dispositional influencings results count. Given this every sensible understanding of the ethics of influencing moral dispositions must commit to at least a moderate sort of instrumentalism. This applies to my own account as well as most competing accounts I'll address (notably in chapters 3 and 5).

By its nature, however, this 'moderate instrumentalism' offers no general insights into the ethical qualities of moral-dispositional influencings. It demands 'influence agents so that they are made better or not worse' or 'influence agents so that more good or less bad is done' but that is as much as it can say in general terms. In each specific situation, of course, we may work out how these demands should best be met based on the character of the situation at hand and the moral-dispositional changes at issue, but in doing this we must (more-or-less) always draw upon ethical resources from beyond what may be usefully called the ethics of moral-dispositional change.

Recall the case of the hitman from the beginning of this section. Someone is influenced to become a hitman (that is, develop moral dispositions such that he's prepared to murder for money) and this is bad. Why? We are not satisfied here by an assertion that such a transformation is just bad-in-itself, absent further explanation. We might instead say, plausibly, that such a transformation is bad because either being a hitman is bad or there being one more hitman around is bad. Why are these things bad? Plausibly, because hitmen kill people and this is a bad thing for them to do and a bad thing for the people being killed and also for those who live in fear of being killed by hitmen and so on. Why are these things bad? Well, we might think human life has a special value worth respecting and that living in fear harms people. Note that the value claims made here (save the initial superficial ones that need to be justified by the latter ones) in justification of an instrumental analysis of why it's bad that somebody be influenced to become a hitman have nothing to connect them particularly to an ethics of influencing. They are instead highly general claims extracted from background understandings of what matters which the instrumentalist applies to the evaluation of the influencing in question.

The overwhelming majority of instrumentalist evaluations of moral-dispositional changes function in this way, by drawing upon claims from some background broader ethics and applying them to the case at hand insofar as they're relevant (the exceptions to this are cases where the results of influencings – as opposed to their conduct – result in violations of the ethics of moral-dispositional

change, such as cases where somebody is influenced to influence in ways deemed bad by such an ethics). This has two major implications for such evaluations.

The first is that it makes any attempt to assemble an 'ethics of moral-dispositional change' from the claims made by such instrumental evaluations of moral-dispositional changes into a very broad project. Such an ethics, were one to be constructed, would reduce to a comprehensive first order ethical theory rather than an ethics *distinctively* of moral-dispositional change of the sort I seek to limit this project to. It would cover all instrumental claims about the ethics of influencing but would also cover so much else besides as to become a different and far larger project.

The second implication, which follows from the first, is that there's little that instrumental evaluations can contribute to the current project. Given that they nearly always depend on such background claims and thus can only provide insight in conjunction with a comprehensive first-order ethics (which is not a reasonable goal for the present project), instrumental evaluations ought to be set aside in the context of a project like this one seeking general insights into just the ethics of influencing moral dispositions.

Think, by analogy, of the ethics of banking. There are many instrumental things that may be said about what a banker should do, given some background set of ethical claims. A banker might discover, say, that his account-holder is an awful person who spends her money supporting awful causes; generating toxic pollution for fun, say. He might recognise that by misadvising and impoverishing her he could do good. Is it right, though, to count the badness of the damage done by pollution as part of the proper ethics of banking? It doesn't seem so. Better to understand this badness (insofar as it is such a thing) as a part of a broader comprehensive ethics (or maybe an environmental ethics, say) that sometimes impacts what bankers should do and try to keep the ethics of banking more specific in content. Similar arguments may be made about the ethics of any human activity or event, from playing sports to changing moral dispositions. While everything is hostage to the evaluative gaze of whatever the best comprehensive ethics is this needn't prevent us from being able to talk about specific areas.

Understanding moderate instrumental evaluation of moral-dispositional influences analogously, I suggest we should develop an ethics of moral-dispositional influencing which may be integrated with instrumental evaluation (and indeed ultimately should be) but which nonetheless forms a *separable module* within ethics. Preparing such an ethics is the goal of this thesis. This approach, I think, offers good prospects for progress in understanding the ethics of influencing moral dispositions by allowing us to narrow focus onto a more manageable subject matter than the totality of first order ethics otherwise necessary for a complete account.

The escape from the need to describe this totality (obviously beyond the scope of the present project) that is analysing a mere module of particularly relevant material rather than this whole does however require an account of how boundaries are to be drawn around this module. It requires an account of how claims integral to 'the ethics of moral-dispositional change' may be demarcated from the 'rest of ethics'.

I propose this may be accomplished through specificity of subject matter. More precisely, I stipulate that a normative claim may only be part of the 'ethics of moral-dispositional change' I investigate if it may more-or-less only be made in evaluating some case of moral-dispositional change.

This rules out, for instance, the claims I mentioned earlier about what's wrong with creating hitmen. These claims seem plausibly (I suggest) to do with the value of human life and freedom from fear and suchlike – things which can impact upon a large variety of moral judgements beyond the

evaluation of influences on moral dispositions. Claiming human life has a certain value, say, implies that we ought to be relatively more careful about risking such lives in circumstances where this is a possibility. It also implies, at least, that in any case where our goals involve sacrificing human lives these goals must be especially strongly justified. Claims integral to the ethics of moral-dispositional change – such that said ethics and projects like mine specifically about said ethics must address them – must impact a much narrower variety of cases. This is what I mean when I say they must impact 'more-or-less only' moral-dispositional change cases; they must impact such cases and relatively few cases of other sorts. They must make a difference to thinking about some rights, wrongs, goods and bads of influencing moral dispositions but relatively few other rights, wrongs, goods and bads.

Also ruled out by this standard are claims about cost. Obviously different methods of influencing moral dispositions are differently costly in many ordinary ways (some might cost money to implement, others time, etcetera) and resulting cost considerations ought to bear upon judgements about the use of specific methods of influencing moral dispositions. Nonetheless, because such costs don't arise only in moral-dispositional change cases they will not be factored into the ethics of moral-dispositional change I explore. Many things other than influencing cost time or money and I've nothing interesting to say about how such costs should be weighed. This isn't to say costs shouldn't be considered, of course, just that they're best bracketed in the context of the kind of project I'm undertaking, at least until the results of this project are ready to be integrated with the rest of ethics (an integration I'll attempt, in a preliminary way, in Chapter 4).

True, certain abstract claims grounded in claims external to the ethics of moral-dispositional change, by this standard, may be such that they may more-or-less only be made in evaluating cases of moral-dispositional change, and thus prima facie fall within the remit I've defined. The claim 'people should be made better by influencing', for example, is one that we can only make in thinking about evaluating actual or potential influencings and the processes of moral-dispositional change they cause (for it is a claim about all such processes; that they are better when they make people better). Nonetheless any interrogation of this claim must depend entirely upon an investigation of what it is for a person to be better or worse, and thus directly and completely on the sort of comprehensive first order ethical consideration I mean to set aside. Thus, it's necessary to stipulate some additional limitation on my subject matter to remove such abstractions.

Thus, I define my subject as *the construction of an ethics (an 'ethics of moral-dispositional change') from claims such that they may more-or-less only be made about cases of moral-dispositional change and which are not such that they depend entirely upon claims about the dispositions being changed themselves* (such as what their consequences might be, what they are, or whether they cohere with others' dispositions or anybody's ideals).

Claims integral to the ethics of moral-dispositional change I pursue might, I grant, depend *somewhat* on the dispositions being changed – on my own positive theory it is true that how one's morals ought to be changed depends partly on the character of the morals being changed. This character just can't provide the whole story, as it might in the case of the hitman (where the bad character of the change is all of what recommends against it) or abstractions like 'influencing should make people better' (where 'better' must be understood entirely in terms of judgements about the character of dispositions).

Of course, we often *have* concerns about the changing of moral dispositions which depend entirely upon the characters of dispositions being lost or gained. We might worry about people being made worse, or bad things happening because of the way influencing leaves people, or even whether

people acquire the same moral dispositions we have. Often in judgement these concerns may be our strongest. This isn't to try to disparage or diminish any such concerns (which have a place in any moderately instrumentalist analysis of the ethics of moral-dispositional change – my own included) – it's simply to say that I'll be concentrating on other concerns. I'm interested in concerns *distinctively about* the changing of moral dispositions, not evaluating dispositions themselves.

This demarcation notwithstanding, there's still a species of content account of the ethics of moral-dispositional change which must be addressed: 'extreme instrumentalism'.

The extreme instrumentalist about the ethics of moral-dispositional change holds that content concerns about the results and characters of moral-dispositional changes shouldn't be set aside to scrutinise other concerns at play precisely because there are no other concerns at play, or at least none worth respecting. In evaluating moral-dispositional changes we should ask whether they make agents better or worse, or make agents such that they do more or less harm or good, or how much they cost, or what benefits they create, but nothing else. Influences upon moral dispositions are ethically uncomplicated tools and ought to be evaluated as such, purely in terms of their results.

One has such 'extreme instrumentalist' thoughts whenever one thinks – as the reader may have – why worry about how moral-dispositional change happens, surely what matters is what it achieves? In extant literature about biomedical moral enhancement in particular Persson and Savulescu have done much to popularise a relatively extreme instrumentalist position on the influencing of moral dispositions. Their book *Unfit for the Future* concentrates heavily on the promised benefits of introducing such technologies, with frequent reference made to the various harms they could prevent and little discussion of non-instrumental concerns about them[84].

The general problem with such 'extreme instrumentalist' views of the ethics of moral-dispositional change is that they reduce to a negative claim about the possibility of any distinct ethics specifically of and distinctively about moral-dispositional influencing. Thus, the truth of any such extreme instrumentalism can only be determined by examining things other than 'content' concerns themselves, such as the 'method' concerns examined in this thesis. Such examinations, if they succeed in discovering any non-instrumental values at play in the ethics of influencing moral dispositions, can disconfirm all extreme instrumentalist understandings of the topic.

This being the case, like more moderate forms of instrumentalism, discussion of extreme instrumentalist understandings of the ethics of moral-dispositional change can be set aside in favour of the 'method' concerns addressed at length in this thesis. While in the former case this approach is adopted for the convenience of being able to work on a more limited set of claims and intuitions, in the latter case this should be done precisely because the soundness of extreme instrumentalism itself is hostage to the outcome of my positive project. Indeed I believe I will show by this project that extreme instrumentalist theories of the ethics of moral-dispositional change fail.

Given all of this, I'll avoid evaluations based on claims about content in this thesis. Instead I'll endeavour to (as far as possible) talk about moral-dispositional changes of general forms which could have a range of actual content. A thought experiment structured around people becoming slowly more charitable, say, will be such that it could equally be structured around them becoming quickly less charitable or committedly pro-life or larcenous. By approaching my subject matter in this way I hope to be able to expose an ethics able to, in general terms, evaluate the means and methods

[84] I.Persson and J.Savulescu (2012), *Unfit for the Future*, Oxford, Oxford University Press; pp.141-143, I.Persson and J.Savulescu (2015), 'Summary of Unfit for the Future', *Journal of Medical Ethics*, vol.41, no.4, pp.338-339; p.338

by which moral-dispositional changes occur than hence the ways in which we become the moral actors that we are.

(2.5) Methods of Moral-dispositional Change

If I'm to evaluate 'methods of moral-dispositional change' (or, more specifically, moral-dispositional change), then I should set out what they are. I can't do this by simply listing them. I could give you examples – 'willpower', certain administrations of oxytocin[85], effective propaganda – but a comprehensive list is beyond my abilities. Worse, giving out a few examples like this and leaving the rest to guesswork is likely to lead to confusion. I treat both the ingenious doctor's disturbing device from §2.2 and innocently reading *Oliver Twist* as 'methods of moral-dispositional change'. I assimilate common-or-garden peer pressures and exotic therapies (if they work) into the same category. Allowing a definition to be 'shown' by the contents of such a varied set is out of the question. It is especially out of the question because, often, I'll discuss such diverse things at once, naming them as part of the same category. I must justify and explain discussing matters in this way.

Firstly, it should be noted that by 'methods of moral-dispositional change' I refer to a class of things including things better called 'means of moral-dispositional change' insofar as they are more like tools than processes. I gloss over this distinction for it is natural here to assess means in terms of their standard methods of use. When I discuss means of moral-dispositional change as methods of moral-dispositional change I do so in this sense; meaning to discuss the means as used in whatever method by which it may be applied to the influencing task in question.

Recall I'm interested in the changing of moral dispositions. Such changes sometimes occur in isolation; the temptations of the smell of sausages may make an ethical vegetarian omnivorous without necessarily making them stop thinking, say, that vegetarianism is a moral obligation. Sometimes, though, they are caused by or cause change in moral ideals. Another vegetarian may become omnivorous by forcing himself to eat meat through willpower having decided vegetarianism is foolish. Still another vegetarian may be tempted to become omnivorous by the smell of meat and then confabulate his vegetarian ideals into oblivion.

Whether changes in moral dispositions are accompanied by changes in moral ideals or not, I'm restricting my focus to the changes in moral dispositions. I leave the evaluation of methods of changing (the evaluation of ways of thinking about how one should act) moral ideals to moral epistemologists.

Often, when moral dispositions change, there's clearly some cause for the change. The vegetarian becomes omnivorous as a result of exposure to the smell of sausages, the cruel man becomes kinder on account of the neurophysiological effects of a drug[86], and so on. Very often, when people's moral dispositions undergo change, we seem able to fit said change into a causal story. Sometimes, plausibly, these causal stories are true; they succeed in describing real causal processes that end in changed moral dispositions.

Sometimes nobody is responsible for these causal processes. An accidental blow to one's head (delivered, say, by a wind-blown branch) might destroy brain tissues that service a habit of charitability (a moral disposition), and thereby eliminate said habit, but nobody could reasonably be held responsible for such an accident. Often, though, these causal processes are such that people do

---

[85] V.Rakić (2017), 'Compulsory Administration of Oxytocin does not Result in Genuine Moral Enhancement', *Medicine, Health Care and Philosophy*, vol.20, no.3, pp.291-297; pp.293-294

[86] *ibid*. p.294

hold some responsibility for them. If you had prior knowledge of such a blow and the changes it would produce, you could have taken steps to avoid it. If some third party had this knowledge, they could have tried to stop the blow or shifted some responsibility by warning you.

It is once agents – agents who expect to have their moral dispositions changed or third parties – become responsible (whether they want to or not) for such causal processes that these processes cease being random events in the world and become 'methods of moral-dispositional change'. A method isn't a random accident; it's something recruited to effect an outcome. Were a third party to become aware of a blow that would break a habit of charitability, and decide it better that it connect and allow it to do so, they would be employing a method of moral-dispositional change against the recipient of the blow. Were the recipient to achieve the same awareness and make the same decision, they would be applying the same method of moral-dispositional change to themselves (this example may be rewritten in terms of inflicting a blow for those concerned about the act-omission distinction).

This, then, is what I call a 'method of moral-dispositional change': a causal process, under the stewardship of some agent(s), with the capacity to effect a change in some agent's moral dispositions. This causal process cannot at any point be wholly mediated by changes in whatever parts of an agent's psychology constitute their moral ideals (although it can be initiated by them). When a method of moral-dispositional change is used by an agent on an agent (who may be themselves) I'll call this an action, though with some methods (such as the use of willpower) this characterisation may be somewhat ill-fitting.

A 'more effective' method of moral-dispositional change is, for my purposes, one more likely to cause expected change(s) in moral dispositions, for we live in a chancy world where not all our schemes go as planned. A truly 'effective' method of moral-dispositional change, a theoretical convenience, is a method certain to bring about the change(s) in moral dispositions that it promises.

Conceptualising methods of moral-dispositional change in this way leads me to ignore prima facie important distinctions of kind. For example, I claim that when a propagandist bombards you with association-rich radio ads[87], so as to (say) make you support some policy, he's doing something relevantly of the same sort as what an imagined mad scientist does who skilfully electrocutes specific parts of your brain in order to achieve a similar outcome. As far as I'm concerned both characters are engaged in the business of moral-dispositional manipulation, they're merely using different methods. They're doing the same thing in the sense that the man who nails planks together with a hammer is doing the same thing as the man who nails planks together by pushing in nails with his index finger. Different more-or-less effective tools are being used to complete the same kind of job.

It could be argued that in committing to this idea I've ignored an important intuitive distinction. Surely, you might say, some means of effecting change in moral dispositions have morally important differences from others, differences that matter in thinking about the ethics of moral-dispositional change? Surely there's something in principle more problematic about the mad scientist than the propagandist? The mad scientist is inflicting changes on your physical body calculated to manipulate your behaviour; the propagandist merely transmits sound over the radio. The former is nightmarish; the latter is annoying. Surely these two characters have done things *different in kind*, and different in kind in a way I can't ignore?

[87] *see* M.A.Bhatt (2012), 'Evaluations and Associations: A Neural-network Model of Advertising and Consumer Choice', *Journal of Economic Behaviour and Organisation*, vol.82, no.1, pp.236-255

This is, of course, the case, but what is at issue is not whether such methods are different kinds of things in all sorts of ways but whether they are the same kind of thing in at least one evaluatively significant way. We might call one 'a performance' and the other 'a medical procedure', but these aren't the kind of kinds that matter in our evaluations of the actions *only as methods of moral-dispositional change*. They matter when we evaluate the actions by the lights of the ethics of advertising, or medicine (or conduct in general), but not by the lights of an ethics governing only processes of influencing moral dispositions. If there's any such ethics (as I think there is), an 'ethics of moral-dispositional change', then there will be many questions specific to thinking about the rights and wrongs of advertising or doing medicine that will be irrelevant to it (as well as some, perhaps, that will be relevant to or parasitic upon it). Questions about ethnic diversity in advertising, say, and questions about the acceptability of taking risks with patients' general health. *Such questions matter* and deserve places in our judgement, including some judgements involving moral-dispositional influencing, but they have no place in thinking *generally* about the ethics of only moral-dispositional change. This much I think I showed in bracketing such moderate instrumental evaluation in §2.4.

This being the case my assimilation of such prima facie distinct actions as propagandising and disposition-altering brain electrocuting into a common evaluative category makes sense. Whatever else they may be and do, both things are methods of moral-dispositional change and however else they should be assessed they should at least be assessed as such. This applies more generally to all methods by which we might attempt to alter the moral dispositions of ourselves and others; each such method should at least be evaluated by any standards that ought to be used to evaluate such things, whatever else ought to be said about it.

(2.6) Personal and Interpersonal Concerns

There are two very different sets of concerns that we might have about methods of changing our moral dispositions. The first set comprises concerns about influence exerted over the moral dispositions of at least one agent by at least one other agent. Classically this is the concern of the subject of propaganda about the actions of the propagandist[88]; it is a concern about things that people do to each other. The second set comprises general concerns about the changing of moral dispositions as can happen by individual choice or accident and which need have no necessary interpersonal dimension. When people debate whether they should take a drug that'll make them act better or take a job that they suspect will make them callous[89], they encounter such concerns.

Both sets of concerns can, prima facie, properly be about methods of changing moral dispositions. In the former case they can have this subject insofar as they can worry that certain methods of changing moral dispositions create bad or wrong power relations when used. In the latter case, they can have this subject insofar as certain methods may be considered bad or wrong due to how they work. Let's call the first set of concerns 'interpersonal' and the second set of concerns 'intrapersonal' (for, unlike interpersonal concerns, they can be such). Both sets of concerns must be analysed in examining the ethics of influencing moral dispositions.

In aiming for some theory of the ethics of moral-dispositional change the best place to begin is with intrapersonal concerns, for if any such concerns have merit (and I will argue some do) they will

[88] *see for example* H.C.Brown (1929), 'Advertising and Propaganda: A Study in the Ethics of Social Control', *International Journal of Ethics*, vol.40, no.1, pp.39-55

[89] *see for example* T.Douglas (2013), 'Moral Enhancement via Direct Emotion Modulation: A Reply to John Harris', *Bioethics*, vol.27, no.3, pp.160-168

impact upon our analysis of interpersonal cases as well as intrapersonal ones. This being the case I defer discussion of interpersonal concerns to Chapter 5, in which I'll argue such concerns ground practical claims coherent with those of the ethics proposed in Chapter 3.

(2.7) Methodology

Now that I've explained what I'll be looking at, I'll declare how I'll look.

My chief goal in this project is to articulate an 'ethics of moral-dispositional change'; a small moral theory describing the proper evaluation of influences on moral dispositions. This should comprise a 'module' of justified and consistent moral claims to be integrated with the demands of comprehensive moral theories to support the evaluation of influences. In seeking this theory, I'll ask myself a number of questions; questions like 'what methods of moral-dispositional change do we find unacceptable?', 'what counts in favour of a method of moral-dispositional change's acceptability?' and 'how do we react to thought experiments about moral-dispositional changes?'. Intuitive responses to these questions, similar questions and questions involved in answering these questions constitutes my dataset. These intuitions can be my own or the communicated intuitions of others. In this thesis I will try to reconcile as many of them as possible into a consistent theory, assuming that these intuitions are not generally foolish or misleading and are hence worth vindicating (at least sometimes). Most of my thesis will work through this process of organising intuitions into an argument for a certain theory of the ethics of moral-dispositional change.

I take these methods to be conventional, the application of an ordinary way of reasoning[90]. Much of this very chapter employs such methods. These methods aren't entirely uncontroversial; Norcross argues they encourage ad-hocery, for example[91]. Nonetheless, detailed discussion of the merits of these methods being the business of a different project, I'll adopt them without further comment.

(2.8) Roundup

This chapter has been an exercise in laying foundations. It has carefully defined my subject matter as the study of the ethics distinctive to changing moral dispositions, a project distinct from investigating influencing more generally (including of non-moral things), moral epistemology and what moral dispositions we ought to have. This done, I should now be prepared to make proper progress towards an ethics of the area thus described.

---

[90] M.Huemer (2005), *Ethical Intuitionism*, New York, Palgrave Macmillan; p.101

[91] A.Norcross (2008), 'Off Her Trolley? Frances Kamm and the Metaphysics of Morality', *Utilitas*, vol.20, no.1, pp.65-80; pp.74-78

Chapter 3: The Rationalizability Requirement

(3.1) Abstract

This chapter offers a thought experiment which contrasts two influencings which employ vacuous rhetoric and accurate news reports respectively. I claim that the former influencing is less intuitively acceptable but that it's hard to say why. Attempts to explain this intuition based on transparency, praxis and non-resistance are assessed and rejected. Instead, I propose that a kind of perceptual and causal connection to relevant realities, comparable to the one identified by Nozick with his 'experience machine' cases[92], is valued in processes of moral-dispositional change. This valuing, I suggest, grounds a requirement on the changing of moral dispositions analogous to one introduced by Brandt with his 'cognitive psychotherapy[93]'. This requirement implies that a more acceptable sort of moral-dispositional changes be 'rationalizable'; such that they work through reaction to presented or represented reasons perceived by changed agents. I hence argue that for any influence N to (with respect to that part of ethics distinctive to changing moral dispositions) legitimately cause some moral-dispositional change C N must present or represent some reason for C and this content must be responsible for N's effect on C. I'll show that this account resists criticisms of Brandt's analogous position, notably owed to Velleman.

(3.2) Annie and Bernard

Imagine two people, Annie and Bernard. Annie and Bernard are similar but separate and unconnected people who face identical instances of the 'common problem' described in §2.3. They don't 'have' the morals that they think they should have; they don't have them because they lack the dispositions needed to practise them. They both, let's say, give tenths of their income to charity but would rather give fifths. For now, though, they just aren't charitable enough to give fifths and continue giving tenths.

Suppose the moral ideals that make Annie and Bernard want to give fifths are identical. Both Annie and Bernard want to give fifths because they recognise the same reasons or have the same second order desire with the same origins, or something similar.

The only difference between Annie and Bernard is the way in which they go about changing their moral dispositions to match their moral ideals.

Annie views news reports which show, by words, sounds and images, the suffering of those who'll benefit from her increased philanthropy. She encounters, through compelling and accurate representations delivered by the reports she views, real people suffering hardship and in need of help. These representations don't give Annie new reasons; they don't, say, show her anything she doesn't already know (nor need they reassure her of anything). Nonetheless these news reports trigger changes in Annie such that her dispositions change. For example, the news reports may induce feelings of sympathy that lead to Annie becoming more philanthropic, giving the fifth of her income her ideals tell her she should give.

Bernard, by contrast, attends to pure rhetoric extolling him to become a more charitable person. This rhetoric is powerful but makes no contact with the realities that Annie's media diet represents.

---

[92] *see* R.Nozick (1972), *Anarchy, State and Utopia*, New York, Basic Books; pp.42-45

[93] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.11

Indeed, let's say the rhetoric fails to represent any of the facts of the world; it is very compelling but also completely vacuous. It may comprise, for instance, an exercise in pure conversational pressure[94]; a very charismatic crowd of people excitedly and inspiringly chanting 'yay charity!'

Let's say both Annie and Bernard, in their different ways, manage to change their moral dispositions. They make themselves disposed to give fifths of their income to charity. Their respective means of moral-dispositional change, it turns out, were effective. Let's say, indeed, that these methods were *equally effective*, equally predictably delivering identical equally long-lasting, resilient, and so on dispositions. A question remains, though: do Annie and Bernard, in achieving their changes, employ morally equivalent ways of changing their moral dispositions?

Intuitively I propose the answer is no. Prima facie, there seems to be something wrong with Bernard's method that's not wrong with Annie's method. Given the choice and in a similar situation, we'd rather follow Annie's example than Bernard's. Giving council, we'd advise acting like Annie before acting like Bernard (ceteris paribus, at least). Plausibly, there's something *a bit worse* about acting like Bernard. Seemingly – and this is my core intuitive claim – *we think there's something pro tanto bad about allowing one's moral dispositions to shift in response to things that don't represent or present reality*. Not just any bit of reality, either, but specific and *relevant* morally important bits of reality (if Bernard achieved his change by studying flowers with no relevance to his charitability we'd still be perturbed; we'd think him odd and, if not too exotic to judge[95], odd in a bad way). All other things being equal we value moral dispositions being, as Brandt said, 'derived from sensitive reaction with the real world[96]'. Other concerns aside we'd rather dispositional change be inspired, motivated, by experience of relevant bits of reality.

Of course, this isn't all we value; we might happily tolerate Bernard doing what he did for the sake of whatever good comes from his increased charitability (especially if he couldn't have achieved it any other way). It is *something* we value in this case, though, I suggest, and it is this value I'll discuss.

The intuitions I suggest AB arouses also seem to occur in relevantly similar cases.

Annie and Bernard initially seem akratic, such that they fail to live up to their ideals due to what's traditionally called 'weakness of the will[97]'. They initially want to give more money to charity than they do but lack the will to do so, and thus seek to fortify themselves by availing of their respective methods of moral-dispositional change. Such akratic characters are useful for exploring ethics distinctive to changing moral dispositions insofar as they exhibit clear divergence between ideals and dispositions. They can believe they should act a certain way or want to be such that they act a certain way in some moral context, but nonetheless be disposed to act differently, and there's little temptation to confuse change in the latter dispositions for change in the former ideals.

This is the case whether the akrasia involved is ordinary or 'inverse'. To employ the classic example of an inverse akratic, imagine two versions of Huckleberry Finn. The first, $Huck_1$, is the character as Twain claimed to understand him[98] and Arpaly[99] (perhaps Bennett[100], depending how we understand

[94] M.Ridge (2003), 'Non-Cognitivist Pragmatics and Stevenson's 'Do so as well!', *Canadian Journal of Philosophy*, vol.33, no.4, pp.563-574; p.564
[95] B.Williams (1985[1993]), *Ethics and the Limits of Philosophy*, London, HarperCollins; pp.162-164
[96] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.157
[97] R.Holton (2003), 'How is Weakness of the Will Possible?' in S.Stroud and C.Tappolet [eds.] (2003), *Weakness of the Will and Practical Irrationality*, Oxford, Oxford University Press; pp.39-67; p.39
[98] L.Bollinger (2002), 'Say it, Jim: the Morality of Connection in Adventures of Huckleberry Finn', *College Literature*, vol.29, no.1, pp.32-52; p.37
[99] N.Arpaly (2002), 'Moral Worth', *The Journal of Philosophy*, vol.99, no.5, pp223-245; pp.228-230

the 'sympathies' he celebrates) interpreted him. This character befriends the slave Jim and helps him to escape, despite believing he shouldn't do so (on account of the antebellum ideals that constitute his 'deformed conscience[101]'), having perceived and been moved by his friend's obvious humanity and oppressed state[102]. The second, $Huck_2$, follows an alternative path. Like $Huck_1$ $Huck_2$ helps Jim escape and like $Huck_1$ $Huck_2$ does this despite believing he shouldn't do so due to the same antebellum ideals that $Huck_1$ accepts. However, unlike $Huck_1$, $Huck_2$'s akratic action isn't born from sympathy for his friend's oppressed condition. Perhaps $Huck_2$ is less attentive to Jim than $Huck_1$ is and is thus less conscious of his plight. However, unlike $Huck_1$, $Huck_2$ wanders into an abolitionist revival. Being Huckleberry Finn $Huck_2$ arrives late and misses all the speeches (they'd bore him anyway) and any chance to be persuaded of the wrongness of slavery by the many reasons they propose. He is however very moved by the revival's closing chant of 'Down with slavery!', which goes on for many minutes with rapturous musical accompaniment and great excitement. It is, indeed, his experience of this chant that renders $Huck_2$, like $Huck_1$, disposed to help Jim escape despite believing he shouldn't.

Obviously, we're glad (sui generis) that $Huck_2$ helps Jim escape however he was motivated to do so, and we might condemn both Hucks for believing they shouldn't help Jim. However, I propose it also seems that there's something somehow better about $Huck_1$'s development of dispositions towards helping Jim than $Huck_2$'s development of (let's say) functionally identical dispositions. $Huck_1$'s being motivated by his perception of his friend's plight seems somehow better than $Huck_2$ being motivated by the revival's chant. This tallies well with my interpretation of AB, which suggests that something about the sort of vacuous rhetoric encountered by $Huck_2$ and Bernard renders it a worse source of inspiration to change moral dispositions, *even when this is all one is doing*[103]. Something about such rhetoric seems to sap it of relevance in a way that the things experienced by Annie and $Huck_1$ are not similarly sapped; something that seems to somehow enhance the evaluative qualities of Annie and $Huck_1$'s respective influencings.

This intuitive response isn't necessarily limited to akrasia cases. We might imagine a strong-willed character, Kim_1, who has the same experiences as $Huck_1$ but who lacks $Huck_1$'s antebellum ideals; instead she's utterly indifferent to the institution of slavery and hence shows no weakness of will in helping Jim having been moved by his plight (though not necessarily having formed any ideals implying that she ought to help Jim). We might equally imagine a strong-willed Kim_2, almost identical to $Huck_2$, who like Kim_1 is indifferent to slavery and thus exhibits no weakness of will when she's moved to help Jim escape by witnessing the closing chant of an abolitionist revival in the same way as $Huck_2$. In considering this variation, I suggest we'll find Kim_1's dispositional change somehow more appealing than Kim_2's, as it was with $Huck_1$ and $Huck_2$ as well as Annie and Bernard.

This sort of intuitive response, which seems to prefer certain methods of changing moral dispositions even when results and disposition-changers' ideals (and these ideals' relations with the changing of dispositions) are held constant[104] also needn't only oppose influencing by vacuous rhetoric. A further example: imagine now Kim_3. Like Kim_1 and Kim_2 Kim_3 is indifferent to slavery and disposed to neither help nor hinder Jim, that is until Kim_3 – by chance – is hit by a cosmic ray from a distant supernova. Kim_3 isn't injured by this, however by extraordinary coincidence the ray strikes certain parts of Kim_3's brain which determine her dispositions towards helping Jim. It flips the charge

---

[100] J.Bennett (1974), 'The Conscience of Huckleberry Finn', *Philosophy*, vol.49, no.188, pp.123-134; pp.125-127

[101] L.Bollinger (2002), 'Say it, Jim: the Morality of Connection in Adventures of Huckleberry Finn', *College Literature*, vol.29, no.1, pp.32-52; p.37

[102] N.Arpaly (2002), 'Moral Worth', *The Journal of Philosophy*, vol.99, no.5, pp223-245; pp.228-229

[103] One is not, say, also or instead shifting one's ideals; one's position on how one ought to act as distinct from how one acts or is disposed to act.

[104] And there exists no external power manipulating disposition changers in any intentional way.

of certain atoms, alters the behaviour of certain neurons. Having previously not been inclined towards helping Jim, following this incident Kim_3 finds herself disposed to help Jim, unable to reveal his location to slavecatchers, say (though her view of whether she ought to help Jim needn't be different), much like the other characters. I suggest there's something bad about Kim_3's moral dispositions with respect to helping Jim being changed in this way, though we might prefer that this occurs all-things-considered (because this helps Jim go free, or Kim_3 have better dispositions). At least, there's something somehow intuitively bad about this change in dispositions that isn't bad about Kim_1's change, or $Huck_1$'s, or Annie's. I suggest that what's bad about this change, in comparison to those others, is that the cosmic ray that triggered it wasn't relevant to Kim_3's moral-dispositional change in the way that the realities perceived by Kim_1, $Huck_1$ and Annie were.

My claim here – then – is that such cases show there to be some intuitive merit to the kinds of dispositional changes encountered by characters like Kim_1, $Huck_1$ and Annie, lacked by the kinds of dispositional changes encountered by characters like Kim_2, Kim_3, $Huck_2$ and Bernard. Setting other considerations (such as results) aside, intuitively the former kind of changes seem better and the latter kind of changes seem worse. I further claim that these kinds are distinguished by the presence or lack of relevant realities inspiring changing persons to change as they do. Taking AB as a characteristic exposure of the intuitions highlighted by these cases I'll attempt, by identifying the considerations that drive our reactions to AB, to make more sense of such cases and hence make progress in understanding the ethics of moral-dispositional change.

One might fret that there seem to be counterexamples I'm neglecting here, however. Particularly, one could point to 'moral fictions' like Dickens' novels and suggest that there doesn't seem to be anything wrong with people being moved by these fictions – which don't seem to be relevant realities – to change moral dispositions. As Cowley notes it's 'no embarrassment' to the social reformer to admit their reformism's literary origins[105]. It would seem there should be such embarrassment, were the intuitions highlighted by AB widespread and genuine.

Such counterexamples don't work. Indeed, close examination of such cases supports my understanding of AB. Dickens' novels, say, famously and systematically, represented real conditions albeit conditions experienced by imagined characters[106]. Dickens was a journalist before writing fiction, and never entirely abandoned his notebook[107]. Although his novels were fictitious, they do many of the things that Annie's news reports do in AB, or $Huck_1$'s and Kim_1's experiences of Jim did for them. They represent the real suffering of real people and the real problems behind this suffering, using the experiences of fictitious people as metaphors. Given this I don't believe the acceptability of Dickens' novels as sources of moral inspiration ('to be inspired', for my purposes, means to be impacted such that one changes dispositions; I also sometimes speak of 'being moved' in this same sense) generates any problem for my intuitive claim.

Moreover, it seems moral fictions generally are acceptable moral inspirations largely insofar as they represent realities relevant to the dispositions they effect.

---

[105] C.Cowley (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290; p.282

[106] T.Sasaki (2011), 'Major Twentieth-century Critical Responses' in S.Ledger and H.Furneaux (2011) *Charles Dickens in Context*, Cambridge, Cambridge University Press, pp.51-58; p.52, 54

[107] J.Bowen (2011), 'The Life of Dickens I: before Ellen ternan' in S.Ledger and H.Furneaux (2011) *Charles Dickens in Context*, Cambridge, Cambridge University Press, pp.3-10; p.8

Think of Stowe's *Uncle Tom's Cabin*[108]. At publication, the abolitionist fiction attracted much anti-abolitionist criticism. This criticism, and subsequent defences, traded largely on whether the book accurately and hence successfully represented the conditions of slavery in then-contemporary America[109]. If the book was accurate, as supporters argued, then the book was a fitting inspiration for one to emancipate slaves, flee slavery, and vote abolitionist. If it wasn't, as anti-abolitionists claimed, then the implication was that the book wasn't a fitting inspiration to do these things. Crucially, the question of whether one should allow the book to inspire you (and it would do so in either case – falsity doesn't, alone, make words uninspiring) for both sides depended on whether the book represented real abuses, real things of moral importance. The question hinged on whether the book was vacuous. This is what we'd expect if my intuitive claim was true – if we really find something wrong with changes in moral dispositions caused by things other than exposure to relevant realities.

One might complain that this interpretation mistakes epistemic debate for debate about the legitimacy of *Uncle Tom's Cabin* as an influence on behaviour. It's certainly true that the book could have changed beliefs (even moral ideals) concerning slavery. The epistemic merits of this changing would have depended on the book's accuracy. Plausibly some debate about the book's accuracy should be understood in such terms, as being about whether the book might be a defective aid in changing beliefs. However, it would be wrong to understand all debate about the book's accuracy in this way. By Stowe's own reckoning, what angered anti-abolitionists about *Uncle Tom's Cabin* was not merely some potential to inculcate defective belief (if it had such potential) *but what it might influence people to do*[110], including people already opposed to slavery[111]. Complaints about this influence combined comment about the effects the book might have on peoples' (notably abolitionists'[112]) behaviour with accusations of inaccuracy[113], trying to use the latter to de-legitimate the former. Such de-legitimation could be attempted because even in 'preaching to the converted' (aiming not or not just to sway belief but to manage behaviour) one doesn't seem liberated from the responsibility to preach only what's relevant and real. More generally, moral fictions aren't merely used to sway opinion but to fortify motivation and inspire. To understand debate about them as purely epistemic is to neglect this.

Moral fictions, then, provide no counterexample to my intuitive claim. Properly understood, they help make it. We seemingly tolerate their influence on our moral dispositions (at least partly) in virtue of their success in representing relevant realities.

Beyond moral fictions there might seem to be other cases where we don't have a problem with people taking moral inspiration from thoroughly vacuous things. It would be wonderful were vacuous rhetoric able to save the addict, say, or change the ways of a psychopathic killer. What such cases have in common is a good thing to be got (or bad thing to be stopped) through the effects of the vacuous rhetoric that's valuable enough to override whatever concerns about proper means we may or may not have – *an overriding instrumental concern*. When an addict requires saving from self-destruction, say, we plausibly have more important things to worry about than doing so in the

---

[108] H.B.Stowe (1852[2016]), *Uncle Tom's Cabin*, Salt Lake City, Project Gutenburg Literary Archive Foundation, available at http://www.gutenberg.org/files/203/203-h/203-h.htm [accessed 17/11/2017]

[109] *see* C.S.Watson (1976), 'Simm's Review of Uncle tom's Cabin', *American literature*, vol.48, no.3, pp.365-368; pp.367-368

[110] C.Parfait (2016), *The Publishing History of Uncle Tom's Cabin*, London, Routledge; p.96

[111] *ibid*. p.96, R.Cavendish (2001), 'Publication of Uncle Tom's Cabin', *London: History Today*, vol.51, no.6, p.54; p.54

[112] T.C.Hagood (2012), '"Oh what a slanderous book": Reading Uncle Tom's Cabin in the Antebellum South', *Southern Quarterly*, vol.49, no.4, pp.71-93; pp.73-74

[113] *ibid*. p.71, pp.73-74

best of ways. We thus bracket some concerns and do whatever works. This doesn't show that in such circumstances means don't matter. Rather it shows that in making judgements about such cases we set aside sufficiently lesser concerns about means just as we usually set aside sufficiently lesser concerns in making judgements generally (I'll discuss how this 'setting aside' should and shouldn't be conducted in Chapter 4).

Given, then, that there seems to be something to our intuitions about AB we must ask what can best explain these intuitions. Before offering my favoured explanation of our intuitions about AB, I'll first go through explanations available in extant literature and endeavour to show why they can't work.

(3.3) Implausible Explanations of AB

Some explanations of our intuitions about AB have less hope of working than others. It is to these implausible explanations that I first turn, to clear the way for more promising options.

Some implausible explanations are moral-epistemic. According to this sort of explanation Bernard is like the fantasist, who forms beliefs and behaviours without reference to relevant realities, and who consequently falls into epistemic error. As epistemic agents our beliefs and behaviours ought to respond to relevant bits of reality and plausibly as moral-epistemic agents we ought to develop morals which are checked by facts like those Annie is exposed to by her news reports and Bernard isn't (at least while enacting his method of moral-dispositional change). Insofar as Bernard is changed by his media diet as Annie is changed by hers he might be lucky but nonetheless epistemically deficient[114]. Moral-epistemic explanations of AB suggest that this deficiency is what the relative 'dodginess' of Bernard's method consists in; epistemic error.

This way of explaining AB fails. It fails as Annie and Bernard are in the same moral-epistemic position. Annie and Bernard share an epistemic position insofar as, ex hypothesi, the things that make them want to start giving fifths of their incomes to charity are identical. Annie and Bernard don't merely share moral ideals, these ideals share justification. They're composed of identical reasons, or prejudices, or confusions, and are epistemologically equivalent. Annie and Bernard may have both read the same article advocating philanthropy or may both been bullied into idealising fifth-giving by the same cult. Either way, insofar as any plausible moral epistemology must judge identical cases identically it can't justify intuitively distinguishing Annie's case from Bernard's.

Recall §2.3's claim that we may pass moral-epistemic judgement on the processes by which we change moral ideals, not the processes by which we change moral dispositions. Moral epistemologies tell us how we should discover what we should do or how we should live[115]. Thus, they give us norms by which to shape change in our moral ideals, those things that codify for us what we should do and how we should live. Moral-epistemic norms can't tell us, though, what to do once we've chosen a new way of acting or living; they tell us how we should make the choice. Thus, it's possible for Annie and Bernard to be in the same moral-epistemic position insofar as they idealise the same morals for the same reasons, despite their different methods for getting themselves to practise them. Given this, it's not clear how invoking moral-epistemic norms might help make sense of intuitions about AB.

---

[114] J.Stone (2013), ''Unlucky' Gettier Cases', *Pacific Philosophical Quarterly*, vol.94, no.3, pp421-430; p.421

[115] A.M.Jaggar and T.W.Tobin (2013), 'Situating Moral Justification: Rethinking the Mission of Moral Epistemology', *Metaphilosophy*, vol.44, no.4, pp.383-408; p.385

Analysing AB as involving Annie acquiring additional reasons from the news reports (like John Harris discussing a similar case owed to Tom Douglas[116]) that Bernard lacks – hence attaining a better moral-epistemic position – can't help here. This is because Annie's reasons for wanting to be someone who donates more to charity (whatever they are) remain fixed in AB being ex hypothesi identical to Bernard's, and Bernard certainly gets no more such reasons from perceiving vacuous rhetoric. In practice, this could be because Annie's news reports don't show her new information, they simply vividly show her things she already knows. Alternately this could be because Annie mistrusts the news reports and resists adjusting her moral ideals in response to them, even as they compel her to give more to charity (making her, perhaps, 'inversely akratic[117]'). Maybe Annie just absent-mindedly doesn't update her ideals, given the news reports. Whatever the reason, per AB's stipulations, Annie's moral-epistemic position must remain constant for it is tied to the ideals she shares with Bernard, even though she's influenced by things which might otherwise improve somebody's moral-epistemic position. Provided Annie's moral-epistemic position remains constant we cannot explain the lesser 'dodginess' of her methods moral-epistemically by citing extra reasons she acquires and Bernard doesn't.

A second set of implausible explanations of AB employ interpersonal explanations which, while they may help elsewhere, don't help with AB. I've defined methods of moral-dispositional change as agent-stewarded causal processes of change in moral dispositions. In AB all the stewarding is being done by the agents subject to the moral-dispositional changes. Annie deploys a method of moral-dispositional change against her own moral dispositions when she elects to consume news reports; Bernard does the same in consuming vacuous rhetoric. No third party is intervening to change either character.

True, somebody else had to create the news reports and the vacuous rhetoric for Annie and Bernard to consume, but the decision to do the consuming was theirs and made knowing the effects these things would have. This being the case it makes little sense to fret about interpersonal matters – such as the influence of one person over another's moral dispositions – when considering AB. No third party is meddling with Annie's or Bernard's dispositions; whatever meddling is going on is *intrapersonal* meddling with one's own dispositions. Any good explanation of AB must respect this.

This is not to claim that interpersonal and broader 'freedom' considerations cannot impact evaluations of influences more generally; they just can't do so in AB (at least in any straightforward way). I'll discuss such considerations in more detail in Chapter 5.

Other implausible explanations of AB are instrumentalist. Perhaps – one might suggest – what distinguishes Annie's methods from Bernard's methods is their respective consequences, for dispositions or the things they impact. The problem with such suggestions is that ex hypothesi Annie and Bernard's different methods of moral-dispositional change have identical consequences. They both yield identical dispositional changes, which are equally predictable, long lasting, resilient, and so on. As such no direct instrumentalist explanation can distinguish Bernard's method from Annie's.

'Indirect instrumentalist' explanations can't tease Annie and Bernard apart either. Such explanations look not just to the results of influences but to the general results of kinds of influencing to make sense of things. They may thus suggest that the kind of influencing Bernard uses tends to produce bad results whereas Annie's methods lack this tendency, and hence try to explain AB. The problem with this response is that my characters' methods seem equivalent in general results as well as

[116] J.Harris (2013), ''Ethics is for the Bad Guys!' Putting the 'Moral' into Moral Enhancement', *Bioethics*, vol.27, no.3, pp.169-173; pp.171-172

[117] N.Arpaly (2002), *Unprincipled Virtue: An Inquiry into Moral Agency*, New York, Oxford University Press, pp.9-11

particular results. Annie and Bernard both use methods able to inspire just about any moral-dispositional changes whatsoever. Bernard could make himself adopt the moral dispositions of a racist, say, by exposing himself to enough racist rhetoric, but Annie could do the same by watching negative news stories about a race she – for some blinkered reason – wants to hate. These stories could be no less true than the reports of suffering Annie uses to enhance her charitability. They might also seem relevant. Their problem is that they give only a limited picture of the world, but this was also the case for the reports of suffering as it would be for any experiential diet save omniscience.

To any who doubt such possibilities, consider that we accept that we may be influenced to become better people in troubling ways[118] (we may be 'brainwashed' into having better dispositions), better people in untroubling ways and worse people in troubling ways. Why reject the possibility of being influenced to become worse people in untroubling ways? Why not admit that sometimes one may be influenced in the right way with bad results?

We might still suppose that perceiving truths tends to inspire better dispositional changes than other origins of such change (or at least perceiving vacuous rhetoric) but this cannot be assumed. Much rhetoric can uplift; many truths (at least perceived in isolation) can mislead. It's a difficult empirical matter whether exposure to truths or rhetoric tends – overall – to yield better results, for any given meaning of better, not something to simply assume. If we must do without assumptions here, though, then it isn't clear how an indirect instrumentalist understanding of AB could render Annie better than Bernard and make sense of the case.

There is, then, no satisfying instrumentalist explanation of Bernard's method's apparent relative dodginess to be found in assessing its relative results, direct or indirect.

(3.4) The Transparency Explanation

Some more promising attempts at explanations of intuitions about AB employ the concept of 'transparency'. One might think the feature of Bernard's method of moral-dispositional change that differentiates it from Annie's is its relative lack of transparency. It's one thing to be moved to help others by observing suffering, it seems quite another to be similarly moved by vacuous rhetoric. The former seems clear and easy to understand, the latter more opaque. Plausibly, prima facie, what we find relatively troubling about Bernard in AB is precisely this opaque quality. Let's call the claim that this sort of difference explains our intuitions about AB – somehow – the 'transparency explanation'.

Christopher Cowley offers a transparency explanation as a solution to the problem of why we tolerate some methods of moral-dispositional change but not others[119].

While prima facie this explanation would seem to engage with moral ideals not dispositions – Cowley speaks of changing one's 'mind', for instance[120] – this is misleading. Cowley characterises the morals he is speaking of explicitly in terms of patterned behaviour, citing the example of a carnivore convinced of vegetarianism's rightness by Singer who continues eating meat and is hence still *morally a carnivore* in the sense Cowley discusses[121].

---

[118] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313

[119] C.Cowley (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290; pp.277-278

[120] *ibid.* p.277

[121] *ibid.* p.282

Cowley suggests what distinguishes 'transparent' from 'opaque' moral-dispositional changes for an agent is the agent's ability to follow what happens during said changes. This introspective task can be achieved for transparent changes but not opaque ones. Cowley argues that the inability to introspect opaque moral-dispositional changes begets suspicion of them and this suspicion in turn explains their troubling-ness and 'illegitimacy'[122]. With reference to AB Cowley would argue that Annie can 'follow' her moral-dispositional change but Bernard cannot, and this difference in turn explains our dissatisfaction with Bernard's methods. Analogous argumentation holding that transparent institutions are more legitimate on account of the capacity of citizens to inspect their workings may be found within political philosophy[123].

What is it for a moral-dispositional change to be 'follow-able'? Cowley describes a carnivore C who is converted to vegetarianism by visiting a slaughterhouse. For Cowley (who accepts Humean projection, though his view needn't depend on the idea) this change will be follow-able, hence legitimate, if C is able to introspect a slaughterhouse-induced change in the morally important qualities of animals as he perceives them such that he will come to judge his remembered previous view of animals (lacking such qualities) mistaken[124]. What must C be able to do to do this? He must be able to recollect his previous state of mind and introspect his present state of mind and recognise the difference and recognise that the slaughterhouse visit caused the change. He must also endorse his present state of mind. C needn't be able to give a detailed explanation of the psycho-mechanics of his moral-dispositional change. C need only be aware of his present and his past views and prefer his present ones.

If only awareness of one's past and present views, and the source of the change between them, plus endorsement, is necessary for 'follow-ability' then Cowley's position faces a problem. The problem is that any case of moral-dispositional change involving an agent with sufficient memory, opinions, introspection and causal perceptiveness will pass the follow-ability test. This stops the position from distinguishing between cases like Annie's and Bernard's. If what counts for transparency is memory, introspection, opinions and causal awareness, why should we suppose that Bernard is any different with respect to these than Annie? Why shouldn't Bernard be able to remember giving tenths, introspect and prefer now giving fifths, and recognise his rhetoric-exposure as what changed him? If there's some reason why we should deny him such things, why shouldn't we also deny them to Annie?

The problem here, and it's a general problem for ethics of moral-dispositional change that base themselves on transparency, is that intuitively troubling and untroubling methods of moral-dispositional change are often equivalently transparent.

This problem endures even when you tighten your account of transparency with additional conditions. For instance, one condition Cowley's account misses out is detailed causal knowledge. You could argue that for a change in moral dispositions to be 'transparent' the agent undergoing said change must be able to describe the causal processes it involves. An account of political transparency, by analogy, ought to prize citizens' knowledge of actual processes of decision-making. A requirement for causal knowledge, though, would still leave you with an account that fails to cohere fully with intuitions. You might know perfectly well what a drug will do to your psychology to make you act better but have no less trouble taking it (knowledge of side-effects aside) than you had

---

[122] *ibid.* p.282

[123] *see* D.Curtin and A.J.Meijer (2006), 'Does Transparency Strengthen Legitimacy?', *Information Polity*, vol.11, no.2, pp.109-122; pp.115-119

[124] C.Cowley (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290; pp.284

when offered it in ignorance of its biochemistry. Annie needn't know more about how news reports impact her psychology than Bernard knows about the effects of rhetoric.

Generally, conditions on a transparency explanation specify extra sorts of knowledge or perceptual access that an agent must have to certify the transparency of their processes of moral-dispositional change. It isn't clear, though, that such knowledge or perceptual access could ever render such processes acceptable. Knowing the details of a wrongdoing, or perceiving it clearly, doesn't usually make it less wrong. To return to the political analogy, transparent corruption is still corruption. If transparent corruption is better than opaque corruption, then this is because it's more easily detected and stopped. The badness of corruption however doesn't issue from its opacity but rather from things like its unjustness or bad consequences. This is why political philosophers don't attempt to evaluate political processes purely using their transparency, as Cowley tries to do with processes of moral-dispositional change[125].

All this makes it strange that Cowley stresses transparency, rather than the things transparency allows us to discern, in making claims about the ethics of moral-dispositional change. Cowley attributes the central place he gives transparency in his account to what may be called the fear or suspicion of the unknown. We – rightly according to Cowley – resist the influence of methods of moral-dispositional change that we lack sufficient introspective knowledge of the workings of. We do so out of our natural suspicion or fear of things we don't know enough about; enough to determine whether the moral-dispositional changes generated by an opaque influence are any good, in relevant cases[126]. They might be 'good' or 'valid' (as Cowley says) but they also might not be good in this or any other sense and qua opaque we can't tell the difference. This inability to tell the difference begets suspicion. This is the suspicion of what the unknown conceals, the fear that it hides dangers.

This kind of suspicion, very generally, sometimes helps and sometimes hinders. The unknown may hold dangers, but also opportunities and things-not-worth-fearing. By presuming it suspect, or fearful, we may miss opportunities or fear what we needn't fear, to the detriment of those impacted by our fear (including ourselves). These generalisations hold for the operation of this fear in the context of assessing methods of moral-dispositional change.

Annie is made a better person (by her own lights) by her consumption of the news reports and made better in a way that seems unproblematic and acceptable. If she was to resist the influence of the reports – avert her eyes and ignore them – out of an ordinary ignorance of quite *how* her behaviours get established, she'd be missing out on a good thing. Indeed, if we are right in thinking (as psychologists sometimes claim[127]) that our knowledge of the formation of our behaviours is often limited, then the consistent application of the transparency view to our judgements in this area risks rendering the domain of moral-dispositional change dark and terrifying, full of scary unknowns that demand resistance. Given this, I think it wrong to give transparency concerns any independent place in thinking about the ethics of moral-dispositional change.

(3.5) Indirect Pragmatism

If transparency differences cannot explain intuitions about AB, then perhaps AB is best explained in terms of differences in something else. Perhaps there are *advantages* that come with moral-dispositional change as done by Annie that don't come with moral-dispositional change as done by

---

[125] *ibid.* p.282

[126] *ibid.* p.282

[127] *see* T.Wheatley (2009), 'Everyday Confabulation' in W.Hirstein [eds.] (2009) *Confabulation*, Oxford, Oxford University Press; pp.203-222; pp.212-216

Bernard. Perhaps what makes the difference is something *practical* tracked by intuitions about AB, something which (as it must, given how AB is specified) doesn't reduce our thinking about AB to extreme instrumentalism.

Allan Gibbard offered such a view in Chapter 12 of *Wise Choices, Apt Feelings*[128]. In this chapter Gibbard, as part of a larger project, outlined standards by which we can judge influences on our norms, including our moral norms and the moral dispositions that sometimes compose them (as 'internalised' norms in Gibbard's model[129]). He captured some of these standards (though not all; he also requires 'plausibility' and in doing so places a content constraint on moral-dispositional changes[130] and commits – as I do – to moderate instrumentalism) in his 'indirect pragmatism'; the idea that we can use pragmatic judgements to work out which influences on norms to accept but do so *indirectly*. The indirectness is needed because, Gibbard argues, once norms are reduced to pragmatic foundations they can lose meaning and stop working. If we reduced norms around grieving purely to pragmatic considerations ('grieve only when it's good for you'), Gibbard suggests, we'd no longer find them worth following[131]. This move distinguishes Gibbard's view from the instrumentalisms discussed in Chapter 2, per which influences are directly evaluated by their results.

Gibbard's indirect pragmatism works by granting legitimacy and relative acceptability to those influences on norms which constitute 'pragmatic considerations'. He means by these 'those things which tended, among our ancestors, toward reproductive success'; those things which broadly promote 'biological fitness'. He argues we must tolerate the influence such things (whether they effect ourselves or others) may have upon our norms, for our extant norms are already the result of their influence – being the products of various fitness-maximising evolutionary processes – and thus if we were to reject this kind of influence we would be forced into a debilitating scepticism[132].

This view of Gibbard's can be applied to AB. It can be applied to AB because it seeks to distinguish between the good methods of moral-dispositional change and the bad ones; to say what distinguishes those influences on norms – including 'internalised norms', some of which are moral dispositions[133] – we should accept[134]. One might propose that Annie's method tends to be more fitness maximising relative to Bernard's. You'll have more or more successful offspring, or live longer, or generally be more biologically fit, if you tend to be moved by news reports rather than vacuous rhetoric. Perhaps our intuitions about AB track this and may hence be captured by Gibbard's indirectly pragmatic model.

Being moved by news reports may or may not tend to maximise fitness better than being moved by vacuous rhetoric. Rather than disputing this proposal I'll offer a direct argument against Gibbard's position and suggest how it may be extended to similar views.

One problem with Gibbard's argument for his position is that it ignores an alternative to scepticism we could employ were we to reject the influence of pragmatic considerations. We could simply sail on Neurath's ship[135]. We needn't, given a disfavoured heritage, immediately reject all of our norms. We could simply go along as before repairing them one at a time, never facing debilitated scepticism

---

[128] A.Gibbard (1990[1992]), *Wise Choices, Apt Feelings*, Cambridge MA, Harvard University Press; pp.219-232
[129] *ibid.* p.71
[130] *ibid*. pp.226-230
[131] *ibid.* p.222
[132] *ibid.* pp.223-224
[133] *ibid.* pp.68-69
[134] *ibid.* p.225
[135] *see* E.Rabosi (2003), 'Some noes on Neurath's ship and Quine's sailors', *Principia*, vol.7, no.1, pp.171-184; p.174

but rather modifying our patterns of influence-acceptance and allowing time and whatever the acceptable influences are to do the rest[136]. If this is possible then Gibbard's justification for indirect pragmatism as a model for evaluating influences trades on a false dichotomy.

The more fundamental problem with Gibbard's story, though, is how counterintuitive it becomes once we unpack what Gibbard means when he says 'pragmatic considerations'. Biological fitness – the capacity to maximise the endurance of one's genes – is a very odd thing for us humans to care about maximising in thinking about the ethics of anything, including the ethics of changing moral dispositions[137]. Granted, it's what evolution 'cares' about (in whatever sense a thing like evolution can 'care' about anything) but we're products of evolution, not evolution. Why should we care about what evolution 'cares' about?

This problem is analogous to a serious problem for 'divine command' theories in ethics, which try to base value upon the goals of a divine creator. Sure, we might be created for some purpose, but why should we care about this purpose unless it's good for some reason that goes beyond it being the goal of our creator[138]? Parents might have a child for the purpose of inheriting the family business, but they don't thereby create a special moral reason for the child. They might create some reason, but even this will be contingent on the child properly owing loyalty to their parents' goals. In responding to such worries Gibbard is arguably worse equipped than the divine command theorist. A god might be supposed a good and authoritative fellow agent – somebody whose council should be respected – something that evolution, a pattern of nature, cannot be. A pattern of nature can't be such a thing because nature contains plenty of good and bad. One can imagine a god however one likes, but mankind has never better personified nature than with a particular Mesoamerican mother goddess depicted committing cannibalism while in labour[139]. Such a thing is too capricious to be worth whatever loyalty is owed to good parents or good and authoritative fellows.

For these reasons I think the Gibbardian approach can't be used to account for intuitions about AB. Even if it can explain AB by somehow showing Annie's method of moral-dispositional change to be on a sounder pragmatic footing it reduces into a dubious Darwinian divine command theory analogue and fails.

Moreover, any approaches sufficiently like Gibbard's will encounter similar pitfalls. The indirectness of Gibbard's indirect pragmatism comes from its employment of biological fitness as a goal for influencing. Were one to instead discipline influencing with one's incidental goals (whatever they happen to be, in the way of ordinary pragmatic judgement) one's position would quickly reduce to an extreme instrumentalism, for one would be left assessing influences entirely according to their service to these incidental goals. Any form of pragmatism meaning to resist this reduction must, like Gibbard's, somehow specify goals for influencing which prevent it from being disciplined by one's incidental goals. In doing so any such pragmatism will risk committing to some counterintuitive background goal defining what it is practical in the relevant sense, as (I've argued) Gibbard does. It will at least be prevented – for fear of reduction to extreme instrumentalism – from judging influences by praxis construed in any ordinary sense. While some future account could navigate this dilemma between reduction and excess, I suggest Gibbard's failure to do so is at least evidence of

---

[136] B.Kelters (2016), 'Gibbard's Indirect Pragmatism: Two Problems', MA Thesis, Sheffield, University of Sheffield; pp.21-22

[137] *ibid*. p.18

[138] *see* W.Morriston (2009), 'What if God Commanded Something Terrible? A Problem for Divine Command Metaethics', *Religious Studies*, vol.45, is.3, pp249-267; pp.249-251

[139] *see* C.F.Klein (2000), 'The Devil and the Skirt: An Iconographic Inquiry into the pre-Hispanic Nature of the Tzitzimime', *Ancient Mesoamerica*, vol.11, no.1, pp.1-26

the difficulty of this task. Enough evidence, I propose, to set aside indirectly pragmatic ways of making sense of AB in favour of other options.

(3.6) Non-resistance

John Christman suggests another approach to evaluating influences, drawing on conceptions of autonomous desire owed to Frankfurt and Dworkin[140] (this 'genetic' tradition of thought about autonomy being distinct from the considerations discussed in Chapter 5[141]). Christman argues that a significant question in evaluating an influence upon somebody's dispositions is whether they, attending to said influence, would resist it. Discussing desires (which, per the definition offered in §2.3, can constitute dispositions plausibly including moral dispositions), he proposes:

> *'(i) A person P is autonomous relative to some desire D if it is the case that P did not resist the development of D when attending to this process of development, or P would not have resisted that development had P attended to the process;*
>
> *(ii) The lack of resistance to the development of D did not take place (or would not have) under the influence of factors that inhibit self-reflection;*
>
> *and*
>
> *(iii) The self-reflection involved in condition (i) is (minimally) rational and involves no self-deception.*[142]'

where to be 'autonomous relative to D' implies that the development of D did not involve influences which were problematic or manipulative in the sense I'm interested in[143]. Indeed, in a recent paper Timothy Aylsworth suggested these criteria can be applied to the kind of evaluative problems highlighted in §1.1[144]; the problems I seek to solve.

Applying these criteria to AB, we might suppose that if Annie and Bernard were to attend to their methods of moral-dispositional self-influencing in the 'minimally rational way' Christman proposes they may come to different views of their respective methods. Annie might consider hers tolerable whereas Bernard might consider his worth resisting.

A distinctive feature of Christman's account is its minimal specification of the evaluation that must confront influences to test their tolerability. Influences must be tested with actual or potential confrontation with subjects' reflective attention[145] however Christman doesn't proscribe the contents of this reflection. While this testing may (perhaps should) involve epistemic criteria where

---

[140] J.Christman (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy*, vol.21, no.1, pp.1-24; pp.5-10

[141] T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; p.694

[142] J.Christman (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy*, vol.21, no.1, pp.1-24; p.11

[143] *ibid.* p.10

[144] T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; pp.694-695

[145] J.Christman (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy*, vol.21, no.1, pp.1-24; p.11

belief changes are involved with causal influences on dispositions this testing reduces to a simple criterion of non-resistance. As Christman puts it 'All that must be true is that the (influenced) agent would not resist – that is, take action to counteract – the process, were she to understand it'[146]. 'Understanding' here requires access to sufficient knowledge of the workings of influences in question[147]. In requiring not just this understanding but non-resistance of these understood workings this position distinguishes itself from transparency accounts. The ability to offer this non-resistance requires only that an agent possess 'minimal rationality'; that is consistency[148], for Christman, and no self-deception[149]. Provided the agent is minimally rational, *whatever influences they can understand but not resist are acceptable*.

The problem with this account is that we can fail to resist understood influences for many reasons, many of which have nothing to do with their character.

Christman gives the example of a 'new cult member' who, two weeks after joining, is left 'mindlessly mouth(ing) the credo of the sect, showing few signs of her former self[150]'. He suggests that, while such a character may retrospectively accept (and reject resisting) the methods used to render her this way, we should see this acceptance as resulting from indoctrination-induced irrationality. We can't, thus, take this acceptance seriously or use it to vindicate the cult's influencings[151].

Suppose, however, that the rejection of resistance involved isn't retrospective in this way. Suppose instead that the cult inductee – call her Linda – accepts the cult's methods prospectively and with the rationality and understanding Christman demands. Suppose that Linda understands that when she joins the cult she'll be subjected to intensive social pressure and sleep deprivation to render her 'mindlessly mouth(ing) the credo of the sect, showing few signs of her former self'. Suppose also, though, that Linda has ulterior motives. Suppose that the cult offers free healthcare to fully indoctrinated members (they check) and their families, and that Linda's daughter needs an operation which Linda can't afford. In such circumstances – perhaps tragically – Linda might rationally and thinking clearly (she needn't be, say, compelled in a way destructive of rationality by concern for her daughter) tolerate the methods employed by the cult. She, like Bernard, simply has goals which overrule concerns about how her dispositions get changed and make her accept influences she might otherwise resist.

I suggest something still goes wrong in this sort of case, however. The reason why Linda accepts the cult's methods in this case has nothing to do with the character of said methods but rather the ends to which they're turned. Indeed, for all we know, the character of the cult's methods in this case might count powerfully against their toleration for Linda, even though she pragmatically decides to accept them. The problem here is that we can accept influences on our dispositions for all sorts of reasons including reasons that have nothing to do with these influences themselves.

Such acceptance may make desires or dispositions resulting from said influences in some sense 'autonomous', given they are chosen in this sort of way. What it doesn't do is address all our concerns about the development of these dispositions (whether-or-not we relate these concerns to autonomy). In Linda's case it seems like something awful was done to her when she was sleep-deprived and pressured by the cult, even if she rationally chose to accept this and her resulting dispositions towards mindlessly mouthing its credo are thus somehow (curiously) constituted from

[146] *ibid.* pp.12-13
[147] *ibid.* p.13
[148] *ibid.* pp.13-16
[149] *ibid.* pp.16-18
[150] *ibid.* p.10
[151] *ibid*. p.19

autonomous desires. This awfulness remains, I suggest, even if we vary Linda's ulterior motive. If Linda was simply seeking company, say, or the protection of tax exemptions given to the indoctrinated, her indoctrination would seem no less awful (though her overall situation might seem less tragic).

Christman's account can't capture such awfulness insofar as it reduces the evaluation of moral-dispositional influences to the question of whether influence-subjects would accept them. Christman argues that there's merit to those influences that we, understanding things fully, would not resist, but this doesn't tell us *why* we should or shouldn't resist particular influences. Given this, Christman's non-resistance account doesn't much advance the evaluation of influences.

Returning to AB, it's possible that in understanding their respective methods of influencing both characters will find them, by their own lights, acceptable. They might both consider fifth-giving important enough to tolerate the method they employ and set aside any misgivings about these methods (like Linda). They may be correct in these considerations (Bernard's methods might seem bad, but him failing to become a fifth-giver might be *worse*). Made aware of each other's approaches they might find Bernard's method less acceptable than Annie's (as I claim). Bernard might even, given this, adopt Annie's method for changing his dispositions (if he can). Citing this relative difference in acceptability in explanation of intuitions about AB, however, is unhelpful. It is to say Bernard should influence himself as Annie influences herself because he thinks that he should influence himself as Annie influences herself (compare, here, the complaint that valorising autonomy – on some constructions – valorises mere continence[152]).

It seems, then, that there are limits to how minimal we can go in the standards we apply to evaluating influences. Mere non-resistance, even reflectively endorsed, isn't enough to helpfully ground such standards. This is because we can engage with influences with ulterior motives which make our responses to these influences a complicated guide to our opinions of them, never mind what these opinions *should be*. Facts about our resistance to influences offer only preliminary insights into our opinions of influences (the kind I reported in making intuitive claims about AB) and not yet explanations of these opinions.

(3.7) A Reality Requirement?

It seems, then, we can't explain AB using transparency, practicality or non-resistance. At least, some strong attempts to non-instrumentally discriminate better from worse influences leveraging these concepts fail. They fail, I suggest, for reasons somewhat generalisable to similar attempts. Transparency alone isn't going to offer us more than an ethics based on the fear of foul play, when what we're really interested in is what makes play foul. Pragmatic approaches are always going to have to employ ordinary notions of practicality and embrace extreme instrumentalism or else risk novel definitions. Facts about whether we resist influences, by themselves, can't tell us much about why we resist them. We're thus no closer to explaining intuitions about AB.

To try to explain these intuitions, I'll now try to say something positive. I'll attempt to outline an explanation of AB that I think has a better chance at working, and show how it may be assimilated into a modification of a framework developed for different purposes by Richard Brandt[153].

Here's my suggestion; what distinguishes Annie's method from Bernard's is the causal connection between reality and Annie's moral dispositions that her method creates. Moreover, it's the

---

[152] *see* A.M.Baxley (2003), 'Does Kantian Virtue Amount to more than Continence?', *The Review of Metaphysics*, vol.56, no.3, pp.559-586; pp.567-572

[153] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; pp.111-129

connection between Annie's moral dispositions and specific bits of reality, 'relevant' ones, that counts and makes Annie's method relatively better. Were Annie to achieve her moral-dispositional change, somehow, by studying flowers, we would find her methods dodgy – if not plain odd – in the same way as Bernard's methods seem dodgy.

This is a big claim. I'm claiming that it matters, somehow, that our moral dispositions result from appropriate causal interaction with specific parts of reality. It's a commonplace that something like this holds for moral ideals – what you think is right very plausibly ought to be influenced by how the world is in some way. It's problematic, though, to apply this thinking to moral dispositions.

It's problematic because in forming our dispositions we respond to particular experiences and events not the general facts we plausibly should respond to in forming our ideals. Thus, as I've noted, an alternative Annie could, say, condition herself into racist dispositions by watching certain selections of even honest news reports, such as ones depicting wrongdoing by members of some race she wants to hate. Such conditioning works through more-or-less direct experience and needn't track how one should act or how one would act all-things-considered[154]. We can and should conquer this limitation by conditioning ourselves in directions disciplined by well-formed moral ideals derived from maximally general facts. The concern that we are conditioned under such discipline, however, instrumentalises influencing; it's a concern that influencings produce good things or better people. When such concerns are temporarily set aside (as they must be for an account of the ethics distinctive to influencing moral dispositions, as demarcated in §2.4) then the alternative Annie and the Annie who becomes a fifth-giver using news reports don't seem that different.

One might propose that per my suggestion – that moral dispositions should result from appropriate connection with relevant realities – we may yet censure racism-seeking alternative Annie for her methods, given the partial nature of the realities she attends to in cultivating dispositional racism (plausibly she couldn't acquire racist dispositions by perceiving all honest news reports; she'd need to attend to specific *partial* selections). There's merit in this proposal. Given my suggestion is that moral-dispositional changes which appropriately connect with relevant realities are better than those which aren't, I can add 'and those connected to *more* or *all* relevant realities are *even better*'. Sensible enough. It is, however, a commonplace that experience is ubiquitously limited and partial. We never see *all* the news reports, nor experience *all* the realities relevant to our behaviour, indeed Sobel argues we cannot do so in principle due to certain experiences being incompatible[155]. Given this there are limits to how much we can reasonably censure influences for only connecting with part of relevant reality, and any attempts to do so must depend on criteria delineating what's too partial – and deserves such censure – from what 'just covers enough'.

It's not clear how one might specify such criteria, however. How should one count relevant realities, as one must if one's to say how many of them are needed for legitimate influencing? How, moreover, might one decide how much partialness in influences' connections to reality to tolerate? One cannot simply say 'if an influence can lead the influenced astray then it must be too partial' for one cannot demand from an ethics of influencing dispositions necessarily improved dispositions at least if results are (temporarily, in keeping with the approach to moderate instrumentalism outlined in §2.4) set aside.

I thus, results aside, don't commit to any line dividing influences which connect with too little relevant reality from those which connect with enough. Rather my view is that appropriately connecting with (representing or presenting, as I will argue, such that said representation or presentation is responsible for the influence's effect) any relevant realities is *always a better-making*

---

[154] *ibid.* pp.98-100

[155] D.Sobel (1994), 'Full Information Accounts of Well-Being', *Ethics*, vol.104, no.4, pp.784-810; pp.801-804

feature in a moral-dispositional influence, whether-or-not said influence is all-things-considered good or acceptable. The more an influence makes such connections the better it is, ceteris paribus (Gibbard's honest civil servant[156] is no counterexample here, for a property can both make something better in the kind of specific sense here discussed and worse sui-generis). Provided an influence appropriately connects with relevant realities at all it attains a certain pro-tanto *legitimacy* (the sense in which I'll always attribute 'legitimacy' to influences in this thesis, unless otherwise specified)[157]. Conversely when a moral-dispositional influence doesn't appropriately connect with any relevant realities it lacks this same legitimacy and we gain some reason to oppose it. None of this claims that people becoming worse or doing bad because of influencing is good or acceptable provided the methods by which influencing occurs are legitimate; we're rather talking about a limited sense of legitimacy here, one which (temporarily) suspends instrumental concerns.

What is it, though, to *appropriately connect* with relevant reality? Such a connection – or something like it – is discussed in one well-known place.

In attempting to refute hedonism, Robert Nozick famously discussed what he called 'experience machines'. These machines simulate for their users a sequence of experiences capturing the users' ideas of the best possible lives. They can simulate the life of a great sports star, successful artist, happy homemaker – whatever you want. They can do it, if you prefer, absenting even the experience of knowing you're in a simulation. Famously, when asked, many people resist the idea of using such a machine. At least, they resist the idea of using it for more than holidays. They resist the lifestyle Nozick proposes, of endless maximally pleasant simulated experience punctuated by breaks to select future simulated experience[158].

Nozick suggests that this tells us we want more from life than happiness[159]. We want something like agency, an ability to achieve something for ourselves. This, though, can't yet exhaust intuitions behind people's reactions to experience machines. It's relatively simple to inject some agency and achievement into experience machines; you need only conceive of them as offering not merely exogenously determined experience but simulated worlds in which you can express yourself and meet challenges, more like contemporary videogames. Even when experience machines are modified in this way, though, our willingness to avail ourselves of them is limited. They still might be fine for holidaying, but there remains something deeply off-putting about the idea of using them permanently or near-permanently. It seems that some part of what motivates dissatisfaction with the experience machine lifestyle is explained by the new and strange relationship with reality they create for their users. To reference *The Matrix*[160], they supplant steaks in people's mouths with the mere stimulation of neural patterns in accord with clever programming. When we taste a steak or admire a scenic vista or kiss a lover, *this just isn't the sort of bit of reality that we want our experiences to be generated by or our actions to consist of*.

Granted, we might be happy to find out that things like the pain of a broken leg are produced by mere machine-stimulation, but only because we'd rather our legs not be broken. When we're insufficiently interested in something not being the case it seems we'd rather our experiences tracked bits of actual reality relevant to their content. Once this preference (which is detectable in empirical investigations of reactions to experience machine cases[161]) is recognised then some

---

[156] *see* A.Gibbard (1990[1992]), *Wise Choices, Apt Feelings*, Cambridge MA, Harvard University Press; pp.20-21
[157] *see ibid.* p.224
[158] *see* R.Nozick (1972), *Anarchy, State and Utopia*, New York, Basic Books; pp.42-45
[159] *ibid.* pp.46-47
[160] *see* L.Wachowski and A.Wachowski (1999), *The Matrix*, Film, Los Angeles, Warner Brothers
[161] F.De Brigard (2010), 'If You Like it, Does it Matter if it's Real?', *Philosophical Psychology*, vol.23, no.1, pp.43-57; pp.46-50

previously obscure debates become sensible; Nozick highlights debate over the use of psychoactives centred on whether they obscure or reveal reality[162]. This preference supplies, indeed, one reason Nozick cites for resisting experience machine use.

The intuition here is shared, I suggest, even by many of those relatively happy to use experience machines. Consider a choice between finding all the nice experiences to be got through the machine in reality itself (suppose one is just lucky) and on the other hand getting the same experiences from the machine. If we didn't care about our experiences being connected with reality in some special way then we would be indifferent between these possibilities – we'd happily settle the matter with a coin toss. We don't feel the need to toss coins here though; rather we'll take the experiences offered by reality over the experiences offered by the machine, when all other things remain equal (unlike in Nozick's original thought-experiment[163]). There's something to reality, it turns out, something that we care about being connected with.

I suspect that this valuing of a connection with reality, ultimately, might result from evolutionary processes like those invoked by Gibbard's indirect pragmatism[164]. In a world like ours where reality holds more risks and opportunities than unreality, there's evolutionary benefit to attending to the former over to the latter. Sometimes it pays not to be lost in one's thoughts, or in somebody else's. This applies to adults anyway, who are responsible for avoiding risks and seizing opportunities; less so to children who lack the skills to do so reliably. Notice, though, that if this sort of evolutionary story explains why we care about being connected with reality it certainly doesn't *justify* the connection. It can't justify it, for the reasons offered in §3.5 for dismissing Gibbard's account of the ethics of moral-dispositional change. It merely gives a just-so story for our intuitions about what's acceptable and what's unacceptable. These intuitions themselves require no further justification; they're just part of the way we humans are. We'll interpret them as justifying; we've evolved that way. If they damn us, in some sense, then they damn us only on account of our humanity.

If we're to take such merely human intuitions seriously, though, being humans, then I think they might help explain AB. Bernard's alienation from reality in AB is of a different sort than the alienation of Nozick's experience machine users. Nonetheless, I suggest this different sort of alienation is of a usefully comparable and important sort.

When Bernard forms moral dispositions, within AB, a certain causal link to the world – more precisely certain important, relevant bits of the world – seems to be absent. His process of moral-dispositional change does involve causal connections to the world, of course – to the rhetorically powerful sounds he listened to – but this doesn't seem to be the *right bit* of the world to cause such a process, just as cleverly programmed circuits aren't for ordinary experiences. The rhetoric doesn't seem sufficiently *relevant* to its effects on Bernard. This is analogous to the experience machine case, where stimulation of neural patterns in accord with clever programming doesn't seem like an okay bit of the world to generate the experience of eating a steak, kissing a lover, or seeing a car.

Bernard's case contrasts with Annie's inasmuch as Annie's process of moral-dispositional change was causally linked to what seems to be one of the right, relevant, bits of the world for the change that it was: the suffering of those who might benefit from her philanthropy. This link in Annie's case, misdirected in Bernard's, was provided by a representational sequence which captured some sort of content from the suffering beneficiaries and presented it to reporters who then represented it in their reports and thence Annie's experience of these reports (this was also the case with $Huck_1$,

---

[162] *see* R.Nozick (1972), *Anarchy, State and Utopia*, New York, Basic Books; pp.43-45

[163] *ibid.* pp.42-43

[164] A.Gibbard (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press; pp.219-232

whose experiences acquainted him with Jim's humanity and oppressed condition[165]). This sort of representational sequence wasn't possible for Bernard insofar as the vacuous rhetoric he attended to ex hypothesi *didn't represent anything*. Consequently Bernard seems to be subject to an alienation from relevant reality that Annie isn't. He changes moral dispositions in response to things that have *nothing to do* with whatever bits of reality are relevant to the changes in question; Annie doesn't.

As with experience machines, then, there seems to be something important that goes missing when the wrong bits of the world cause change in moral dispositions; a connection with relevant reality that Annie maintains and Bernard is alienated from. As Specker et al. note, it seems like some sort of important content drains from moral-dispositional changes in such circumstances[166]. It seems like, in positions like Bernard's, one's dispositions fail to connect with – are alienated from – the relevant realities they should be caused by in a manner that is bad. This, indeed, seems in some ways a worse sort of alienation from reality than that encountered by the experience machine user. Experiences pass, dispositions remain. You can holiday in an experience machine; you can't vacation from your own character and the moral dispositions that help compose it.

For these important dispositions, then, to be caused to change by things other than the realities that ought to cause such change would seem to put the changed person in a bad and tragic state. It's to enter a state where some of the most important features of one, one's moral dispositions, have their origins in happenstance or artifice rather than the bits of reality that ought to shape them. To acceptably enter or put another into such a state would seem to at least require some excuse. It would otherwise seem to violate what we might call, preliminarily, a 'reality requirement' on the ethical changing of moral dispositions; a requirement that (absent sufficient reason to do otherwise) our moral dispositions be changed only by exposure to representations or presentations of relevant realities.

(3.8) Towards a Positive Account: Relevance in Brandt's Cognitive Psychotherapy

Richard Brandt offered a usefully analogous position (with different purposes in mind) in outlining what he called 'cognitive psychotherapy', an approach to properly forming the desires that he believed determine human behaviour[167]. Key to this methodology was the view that the proper way to form these desires is to confront oneself with vivid representations or presentations of relevant bits of reality. For instance, someone averse to a job because of his parents telling him it's beneath him ought to attend to the particulars of doing the job, and in such a way restructure his desires[168]. Someone looking to stop pining over a lost love ought to attend to their love's annoying features[169]. This approach mirrors the preceding reflections on AB and particularly the notion that our moral dispositions ought (ceteris paribus) to be changed by causes which work through presentations or representations which connect us with realities relevant to the changes in question.

Brandt's 'desires' are more specific things than the psychological jumble I've ecumenically called 'moral dispositions' which are, recall, dispositions to act of the sort that agents actually have, constituted by whatever features of agents' psychologies – habits, behaviour-causing desires, motivating concepts, etc. – actually constitute such things. They have complex inter-substitutability

---

[165] N.Arpaly (2002), 'Moral Worth', *The Journal of Philosophy*, vol.99, no.5, pp223-245; pp.229-230

[166] J.Specker, M.H.N.Schermer and P.B.Reiner (2017), 'Public Attitudes Towards Moral Enhancement. Evidence that Means Matter Morally', *Neuroethics*, vol.10, no.3, pp.405-417; p.414

[167] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.113

[168] *ibid.* p.157

[169] C.Rosati (2000), 'Brandt's Notion of Therapeutic Agency', *Ethics*, vol.110, no.4, pp.780-811; pp.784-785

criteria[170], causal connections with pleasure[171] and exhibit certain compatibility relations[172]. Nonetheless Brandt's desires have two features which make at least some of them at least a subset of dispositions in my sense; they're sometimes somewhat stable (or a least arise in stably patterned ways)[173] and they can cause behaviour[174] (Brandt sometimes muddies things by speaking of 'valences', but for him talk of these *is* talk of desires[175]). Given this, and Brandt's model having application to moral contexts[176], when Brandt makes claims about the way desires ought to be changed he also makes claims about how moral dispositions ought to be changed. Not *all* moral dispositions, true, but many of them. This being the case, Brandt's cognitive psychotherapy offers a useful departure point in developing an account of the ethics of influencing moral dispositions, consistent with my understanding of AB, especially insofar as he also addressed the species of 'relevance' invoked in §3.7.

Granted, Brandt had different goals from me in articulating his cognitive psychotherapy. He wanted an account of what it is to be rational[177], not an account of how moral dispositions should be influenced. Given this my engagement with his position *shouldn't be read as an attempt to continue or defeat his metaethical project*, so much as to engage with some of his ideas to inform a different project. Provided said ideas are about the same sort of things, or at least some of the same sort of things, that I discuss – as I argue – resulting claims shouldn't be undermined by this difference in goals. Brandt's views should be able to help me work towards a more detailed positive story.

I've claimed that we value moral-dispositional changes being caused by experience of relevant realities. What, though, must be the case for some part of reality to be 'relevant' to a moral-dispositional change?

Brandt first discussed relevance (which may have different meanings in other contexts, notably in pragmatics[178]) in moral-dispositional influencing while critiquing emotivism and Stevenson particularly[179], later integrating the concept into his own mature position[180]. Criticising Stevenson[181], Brandt argued that his emotivism reduced what Brandt called the 'logical relevance' that we discern between things and moral dispositions (and Brandt and Stevenson *are* discussing moral dispositions here insofar as they generally discuss the relevance of influences to 'ethical attitude[182]', itself explicitly defined dispositionally[183]) to 'causal relevance'; the bare fact that a given thing causes a given moral-dispositional change. This was a problem for Stevenson, Brandt argued, because we can call certain causal influences on moral dispositions 'irrelevant', even if they do cause changes in the moral dispositions they're irrelevant to. On Stevenson's model, Brandt contended, this ordinary way

---

[170] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; pp. 32-35
[171] *ibid.* p.96
[172] *ibid.* p.332
[173] *ibid.* p.82
[174] *ibid.* p.30, pp.48-51
[175] He indexes the former to the latter; *see ibid.* p.362
[176] Which it does; *see ibid.* pp.170-171
[177] D.Sobel (1994), 'Full Information Accounts of Well-Being', *Ethics*, vol.104, no.4, pp.784-810; p.792
[178] *see for example* R.Carston and S.Uchida (1998), *Relevance Theory: Applications and Implications*, Amsterdam, John Benjamins Publishing Company; pp.8-11
[179] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.309-313
[180] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; pp.111-112
[181] *see particularly* C.L.Stevenson (1944), *Ethics and Language*, London, Yale University Press
[182] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, p.312
[183] C.L.Stevenson (1950), 'Brandt's Questions about Emotive Ethics', *The Philosophical Review*, vol.59, no.4, pp.528-534; p.530 (footnote)

of attributing relevance makes no sense; Stevenson claims causal effectiveness sufficient for relevance and this precludes something being both an effective influence and irrelevant to the dispositions it influences[184]. Stevenson's rejoinder was to challenge Brandt to explain what he meant by 'logical relevance', beyond 'a property of influences that are regarded as being such that they ought to make a difference to the moral dispositions they influence[185]'.

Stevenson himself could suggest an understanding based on approval[186], calling 'relevant' those influences which are approved of by attributors, but Brandt thought this problematic because one might approve of an influence but also find it lacking relevance [187]. We might, say, approve of moral-dispositional change induced by an influence for the sake of some greater good but still regard the influence as lacking relevance to what it changes (recall Linda from §3.6). One could, from an emotivist perspective like Stevenson's, analyse such cases as involving distinct sorts of approval, one of which works at a general level and one of which works to attribute relevance by taking into account only some specific kind(s) of considerations. This would however require a more complex and substantive story than Stevenson's[188].

Importantly in debating Stevenson Brandt thereby suggests there's something substantive to 'relevance', as can relate influences to moral dispositions. The present project requires exactly such a substantive 'relevance' to support discriminate claims about which influences are relevant to which dispositions and advance from what I've called the reality requirement to a detailed account. Thus, I'll consider how Brandt met Stevenson's challenge.

Brandt attempted to define relevance as part of his mature metaethics.

As someone who understood an agent's good in terms of what they would desire, if sufficiently well-informed[189], Brandt needed to explain how an agent might become well-informed. This is a standard task of the 'ideal observer' theorist in ethics; having defined good in terms of the ideals or dispositions of some idealised agent, they must then describe such agents or at least how one might become like them[190]. Brandt's idealisation – 'full rationality' (as he called it) – was intended to be more reachable and less impersonal than the omniscience or impartiality (say) required by those he called 'ideal observer theorists[191]', but his mature account[192] nonetheless presented him with a comparable dialectical burden.

To lift this burden Brandt introduced 'cognitive psychotherapy'; a method of making 'rational' one's desires through undoubted and maximally vivid relevant experience[193]. By application of this method, Brandt claimed, one might become more 'rational' such that one desires what is good for

---

[184] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313

[185] C.L.Stevenson (1950), 'Brandt's Questions about Emotive Ethics', *The Philosophical Review*, vol.59, no.4, pp.528-534; p.528

[186] *ibid*. pp.528-529

[187] R.Brandt (1950), 'Stevenson's Defence of the Emotive Theory', *The Philosophical Review*, vol.59, no.4, pp.335-340; p.336

[188] *ibid*. p.336

[189] C.Rosati (2000), 'Brandt's Notion of Therapeutic Agency', *Ethics*, vol.110, no.4, pp.780-811; pp.784-785

[190] T.Jollimore (2020), 'Impartiality' in E.N.Zalta (2021) (ed.) *Stamford Encyclopaedia of Philosophy*, Fall 2021 Edition; *see* §2.2

[191] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.225

[192] *see generally ibid.*

[193] *ibid.* pp.11, 111-112

one[194]. More importantly for the present project, this method required some account of how to curate experience to achieve a proper course of 'cognitive psychotherapy'. It was in this context that Brandt proposed an account of 'relevance', meaning to distinguish between those parts of reality whose perceptions have places in cognitive psychotherapy programs – and hence ought to influence us – from those which do not.

In addition to causal effectiveness[195], Brandt defines relevance using two conditions. The first is a sort of 'veracity condition', and treats 'relevance' as a relation between beliefs and moral dispositions. For a given belief to be relevant to a moral disposition, according to Brandt, it must be 'part of the 'scientific knowledge' of the day, or … justified (by available deductive or inductive methods)[196]'. What Brandt seems to demand here is that relevant influences be verified; withstand the scrutiny of our available epistemic toolkit. Brandt's second condition treats 'relevance' as a relation between influences in general and moral dispositions. Brandt cites the example of using the onerous completion of multiplication tables as a method for extinguishing the desire to drink Martinis by repeatedly doing the two things together[197]. This method involves irrelevancy, for Brandt, insofar as it could be used to diminish any desire whatsoever. This shows the onerous completion of multiplication tables to simply be capable of generating negative associations rather than something *relevant* to the activity of Martini-drinking. Relevancy, Brandt thinks, requires some connection linking an influence and the dispositions it influences. The influence must be 'for instance, a thought about the expectable effects of the thing, or about the kind of thing it is, or about how well one would like it if it happened, and so on[198]'.

More must be said about Brandt's conditions. His 'veracity condition', in particular, may only be applied to beliefs and other suitably belief-like 'representational' things – only things subject to verification. In using Brandt's model to try to make sense of relevance as it can exist between *any* influence on moral dispositions and moral dispositions, this fact creates a problem. If the veracity condition as offered by Brant is left unaltered any non-verifiable influence is necessarily declared irrelevant. This seems too quick insofar as not only verifiable representations of the facts of the world but unverifiable[199] facts of the world themselves sometimes seem relevant to our moral dispositions.

Consider Roger who, confronted with a homeless person freezing in the snow, is deeply moved and immediately develops new moral dispositions to try to help the homeless. Roger needn't learn anything – he needn't acquire new beliefs – from this confrontation. Roger might always have known that the homeless suffer when it snows. Roger might even have known that the particular homeless person before him suffers when it snows. Yet, despite Roger's lack of fresh beliefs, we prima facie find his change in moral dispositions unproblematic[200]. If we've any complaints about

---

[194] C.Rosati (2000), 'Brandt's Notion of Therapeutic Agency', *Ethics*, vol.110, no.4, pp.780-811; pp.784-785

[195] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.12

[196] *ibid.* pp.111-112

[197] *ibid.* p.112

[198] *ibid.* p.112

[199] Facts are factive; they are or are not; it makes no sense to ask whether they are or are not verified in the way it makes sense to ask whether representations of facts are verified. Facts may make it such that certain representations may be verified, and we may metaphorically ask whether facts are verified (when we mean 'is the representation of the facts – say in some given sentence – verified?'), but neither of these things reflects on the verifiability of facts.

[200] R.Noggle (1996), 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly*, vol.33, no.1, pp.43-55; p.49

Roger they won't be that he *was* moved by the sight of the suffering homeless person so much as that he *wasn't* moved previously by his existing belief in this suffering.

Moreover, we don't just find this change unproblematic because we think it's a change for the better. There's something different and more acceptable about Roger's change compared to an alternate character who undergoes an identical dispositional change upon seeing Mount Everest (provided, that is, that seeing Mount Everest itself causes dispositions to change rather than some other more relevant representation or presentation the character is, say, reminded of or encouraged to imagine by the sight). This wouldn't be the case if we cared only for results.

We can accommodate such cases if we regard parts of reality itself as amongst the things properly relevant to moral dispositions. The suffering homeless person in a snowstorm is a bit of reality plausibly relevant to moral dispositions determining treatment of the homeless, just as the persons' beliefs about this suffering represent bits of reality which plausibly should determine this same treatment. Perceiving the former can (other considerations aside) acceptably inspire someone to start treating the homeless differently in a way that perceiving mount Everest cannot and reflecting on true beliefs about the former can (this isn't to say one can't be too sentimental and thus do wrong).

Given this understanding, Brandt's 'veracity condition' subsumes into the reality requirement suggested in §3.7. What fundamentally matters isn't that we're inspired by sound beliefs or other representations *but by bits of reality that ought to inspire us*. Hence, inspiration unmediated by sound belief can sometimes be acceptable.

This isn't a surprising finding with respect to belief provided it's remembered that it (according to all bar those pragmatist approaches which assimilate beliefs into dispositions[201]) is essentially a mediator between reality and dispositions. Beliefs represent reality in ways that allow it to shape our dispositions. As such, it shouldn't be surprising that a condition limiting what beliefs ought to be able to do is explicable in terms of deeper conditions limiting what the reality beliefs represent ought to be able to do. The same goes for all representations composing the totality of veracity-apt things. The explanation of Brandt's veracity condition is that veracity coordinates representations with reality; meeting the veracity condition is good insofar as it helps representations meet the reality requirement. The reality requirement, though, is more fundamental. We ultimately want moral-dispositional changes inspired not by sound beliefs (especially if 'soundness' is constrained by fallible things, like whatever the best available epistemology is[202]), but by things which are the case and ought to inspire us (and these two things must be distinguished[203]).

If Brandt's first condition may be subsumed into the reality requirement then it can't advance us much (though it rightly points out that, with regard to compelling representations, verifiability evidences the meeting of the reality requirement). We don't yet have enough for a full account of relevance. Both the freezing homeless person and Mount Everest from my previous example are parts of reality. If being a part of reality *alone* was sufficient for relevance then neither inspiration would be more acceptable than the other. More is needed here. What's needed is some set of conditions which can explain not just why reality has some special license to effect change in moral dispositions – one which may be transmitted to representations (in a way that verification evidences) – but why certain parts of reality have this licence and not others.

---

[201] C.Hookway (1998),'Doubt: Affective States and the Regulation of Inquiry', *Canadian journal of Inquiry*, vol.28, no.1, pp.203-225; p.206

[202] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; pp.111-112

[203] D.Sobel (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492; p.476

Brandt's second condition offers a good starting place in seeking these additional conditions. This condition demands that a relevant influence on a moral disposition be somehow connected with the moral disposition in question. This is intuitive; relevance is a relation and it's sensible to analyse it in terms of some kind of relatedness[204].

Brandt attempts to define this relatedness in two ways. He first offers an example – the example of reciting multiplication tables to curb one's martini-drinking – where it seems the relatedness is missing. This is shown by the fact that such recitation might effect change in (specifically, diminish) any disposition whatsoever. Brandt then suggests that an influence may only be relevant to a dispositional change if this change is, somehow, 'a function of its content'. He offers the example of how a thought could be relevant to a desire by being 'in some fairly restricted way about the thing desired'. It may be 'about the expectable effects of the thing, or about the kind of thing it is, or about how well one would like it if it happened, and so on[205]'. What this 'fairly restricted way' *is*, however, is not explicated but rather illustrated by these examples of how the necessary 'about-ness' may be achieved.

It's clear that these two attempts at definition aim to pick out a common thing, the further relation that is cited by Brandt's second condition on relevance. Both of these attempts, though, are problematic.

The problem with defining relevance in terms dependant on what in fact effects people's dispositions, as per Brandt's first attempt, is that it seemingly allows for counter-examples.

Brandt cites doing multiplication tables as irrelevant to the disposition to drink martinis on the basis that the activity could diminish any disposition[206]. What if, though, in some moods I enjoy doing multiplication tables; I find it a relaxing mental exercise, say? What if, indeed, I only get into these mathematical moods in certain settings and while doing certain things – only while jogging, say? If I happen to be this way inclined, then for me (at least) it doesn't turn out to be the case that doing multiplication tables can diminish any moral disposition whatsoever; it can't diminish my dispositions towards jogging. Does this show that for me doing multiplication tables is *relevant* to my disposition to drink martinis?

Indeed, I needn't even be such an oddly mathematical person to be such that doing multiplication tables can't diminish in me any disposition whatsoever. I need only have some *deep-seated* disposition somehow entirely insulated from diminution by any influence. If I have any such dispositions then for me no influence will be such that it can diminish any disposition whatsoever.

These problems persist even if we, charitably, allow that Brandt doesn't demand that illegitimate influences effect *any dispositions whatsoever* but just *too many* dispositions. I might have many deep-seated dispositions, such that no influence effects too many of my dispositions. I might get into mathematical moods doing many different things, such that reciting multiplication tables can't effect too many of my dispositions.

Something has gone wrong here.

---

[204] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.112

[205] *ibid.* p.112

[206] *ibid.* p.112

What's gone wrong is that Brandt – like Stevenson, per Brandt's own criticism[207] – has attempted to define relevance causally in terms of the actual effects of influences on dispositions[208]. He does this insofar as he labels irrelevant influences with the causal capacity to alter 'any disposition whatsoever' (or, more charitably, too many dispositions). The problem with doing this is that people have systematically divergent susceptibilities to influences. We don't all have the same psychological features for influencers to leverage nor are these features unchanging[209] nor are there principled limits on their characters. As such for any account of the relevance of influences to dispositions grounded in a given set of actual human susceptibilities to influencing there will always be room for people to diverge from the norm such that the account delivers counterintuitive results.

This isn't to say that the idea that some influences can do too much doesn't have some role in discerning which influences are relevant or legitimate. In *A Clockwork Orange* the capacity of the 'Ludavico method' to make one avoid Beethoven as well as violence seems to count against it[210], even if this capacity isn't at the root of what is wrong with it. I'll later suggest that the capacity to influence many dispositions evidences that a given influence is capable of exerting influence despite irrelevance, though this capacity cannot be used to define irrelevance for it isn't necessarily co-present with irrelevance. It does *tend* to be co-present with it, though.

(3.9) Relevance and Rationalizability

The problems with Brandt's second attempt to define the relevance relation are interpretational. What's the 'content' of a potential influence on a moral disposition and how must it be related to the moral disposition it influences such that it will be relevant to whatever change in it it causes?

Brandt clarifies his definition by suggesting, as a sort of thing with the right sort of content, 'a thought in some fairly restricted way about the thing desired; … about the expectable effects of the thing, or about the kind of thing that it is, or about how well one would like it if it happened, and so on[211]'. Given this, we might think the significant sort of content is being such that the content-holder may be properly called 'about' – might represent[212] (or present, given what I've said about the verifiability condition[213]) – some *reason* for or against the prospective dispositional change.

That a dispositional change has a certain expectable effect, say, in many circumstances, might count in favour of or against making said dispositional change. That giving more to charity will have the expectable effect of reducing suffering, say, seems to quite generally count in favour of becoming disposed to give more to charity. That giving more to charity will have the expectable effect of lowering one's living standards, say, seems to quite generally count against becoming disposed to give more to charity. Similarly, in many circumstances the kind of change that a dispositional change is (whether it involves one becoming more charitable, say) or how well one would like it if one made said change might count in favour of or against making said dispositional change. This is not the case

---

[207] R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313

[208] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.356-357

[209] *see for example* A.Mukhopadhyay and G.V.Johar (2007), 'Tempted or Not? The Effect of Recent Purchase History on Responses to Affective Advertising', *Journal of Consumer Research*, vol.33, no.4, pp.445-453

[210] L.Calhoun (2001), 'At What Price Repentance? Reflections on Kubrick's *A Clockwork Orange*', *Journal of Thought*, vol.36, no.1, pp.17-34; p.26

[211] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.112

[212] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; p.365

[213] I'll often say 'show' to mean 'represent or present'.

for all other features of dispositional changes. That a change happens on a Tuesday, say, seems unlikely to count for or against making said change in nearly all circumstances. I suggest that in picking out, by example, a set of things relatively likely to be reasons for the dispositional changes their perceptions cause, Brandt meant to define the content of relevant influences in terms of the relevant reasons they represent.

This interpretation yields my favoured analysis of 'relevance': 'to be relevant to a dispositional change is to successfully represent (or present, an additional possibility allowed whenever I say 'represent' in defining 'relevance') some reason to make the change'. For any influence, to be relevant to a disposition itself (a shorthand) is to be relevant to some process of change in said disposition (in particular whichever one is being evaluated). The reasons needed to ground relevance (as opposed to, say, the mere appearance of relevance) must be *justifying*; they must indeed count in favour of the dispositional changes influences cause[214], in whatever way that reasons can indeed count in favour of things. The 'content' Brandt's second attempt to define 'relevance' was getting at, on this understanding, is representational or presentational content that might otherwise play some part in[215] rational deliberations about whether the subject of influence should undergo caused dispositional changes. One may suggest calling the class of dispositional changes inspired by such content *rational* (with, roughly, Arpaly[216]) and this may sometimes be apt, however I prefer *rationalizable*, for one might demand from rationality proper some coordination of one's dispositions with one's beliefs or ideals that needn't be present in such changes.

This interpretation is perhaps unsurprising; Brandt was aiming to articulate an account of rationality and perhaps reasons themselves[217]; it's natural that such a project should characterise a key operator like relevance in terms most readily understood with reference to reasons. As Brandt's project, however, had to be prior to any understanding of the nature and distribution of reasons, to avoid question-begging, he couldn't commit to the analysis of relevance I do here. In a sense I'm proposing an understanding which mirrors Brandt's but is reversed; it defines relevance with reference to where reasons are and not where reasons are with reference to relevance. Insofar as I'm undertaking a limited first-order investigation into what we should do (a different project from Brandt's) I can help myself to a background understanding that some things are reasons for some other things without committing to a specific theory of said reasons' natures and distribution and define relevance by reference to this understanding.

At this point though I must add to this definition of relevance that *entire* relevance for an influence does require that the representational content of said influence be *on its own* the influence's influential part.

Imagine somebody who becomes aware that he's becoming disposed to practice a given religion (if not yet believe it) as a result of appreciating its art. He might respond emotionally to its music, say, and suddenly discover himself being drawn into its practice; say into attending its services or observing the moral code it endorses (some religious art is created to impact dispositions in such

---

[214] *see* J.J.Tiley (2004), 'Justifying Reasons, Motivating Reasons and Agent Relativism in Ethics', *Philosophical Studies*, vol.118, no.3, pp.373-399; p.376

[215] Though not necessarily determine the outcome of; the reasons represented or presented in this content need only be pro tanto.

[216] N.Arpaly (2002), *Unprincipled Virtue: An Inquiry into Moral Agency*, New York, Oxford University Press, pp.50-59

[217] D.Sobel (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492; p.476, p.476 (footnote)

ways[218]; suppose some of it sometimes works). This character may think 'okay, so this song does represent some reasons to make the dispositional changes it triggers – it does say how happy I'll be if I live by the religion's teachings – but is it inspirational on account of saying this or because it sounds pleasant?' In asking this the atheist is asking whether the representational content or the aesthetic qualities of the song produce its effects on his dispositions.

This is an important distinction. Any representation, after all, no matter how individually uninspiring, can be associated with strong emotions and given force. *Triumph of the Will* presented an infamously effective spectacle[219], but its Nazi triumphalism wouldn't have moved audiences so much absent the spectacle. The distinction here is between being inspired by relevant reasons, represented or presented to one, and being inspired by the 'stuff around' such representations and presentations – the music they're set to, the setting they're relayed in, the kind of day you're having when you encounter them. The latter things are irrelevant but they nonetheless have the capacity to make a representation more or less able to influence one's dispositions, either by modulating its effects on one or adding effects of their own.

This capacity for appropriately connected irrelevancies to change the power of an influence on moral dispositions is behind, I think, Brandt's second condition; the one demanding specificity of effect[220]. The fact that irrelevancies can add inspirational power in this way to anything makes the fact that a given influence could trigger any dispositional change whatsoever evidence that said influence gets its power from factors not relevant to its effects. Thus 'the capacity to trigger any moral-dispositional change whatsoever' evidences the irrelevance and consequent pro-tanto unacceptability of a means of changing dispositions but, contra Brandt, can't help with defining said irrelevance.

My view, then, is that for an influence N to be *entirely relevant* to a dispositional change C (in the sense required for it to be maximally pro-tanto legitimate with respect to the ethics distinctive to changing moral dispositions, as distinguished in §2.4) two things must be the case. Firstly, N must represent or present some justifying reason to make change C. Given the reflections of the preceding four paragraphs I must secondly add that this representational or presentational content by itself must entirely originate the influence exerted by N.

Granted, the notion of 'entire relevance' here is an idealisation. Most real influences, at best, cause changes in dispositions which partly result from their representation of relevant reasons and partly result from the modulation of this influence, or additional influence, by other features they have (such as certain non-representational features or representation of irrelevant things). The thought here is, though, that influences are more relevant and hence better (ceteris paribus) the more they approach this ideal. All other things being equal, the more an influence uses representation of relevant reasons in to effect whatever changes in moral dispositions it effects the more relevant and acceptable it is. In other words there's what I'll call a 'rationalizability requirement' on ethical (in a certain pro tanto sense, distinctive to evaluating moral-dispositional influences) moral-dispositional influences.

My proposed understanding of relevance, as can relate influences to dispositions and (in a certain pro tanto sense) legitimate the changing of the latter by the former, differs from Brandt's. Unlike Brandt's understanding of relevance, which was meant to be prior to and inform an understanding

---

[218] *see* F.A.Luchs (1950), 'To Religious Experience by Pathways of Art', *Religious Education*, vol.45, no.6, pp.349-352

[219] W.Brown (1997), 'Triumph of the Will', *History Today*, vol.47, no.1, pp.24-28; p.28

of the nature and distribution of reasons[221], the understanding of relevance here proposed is posterior to and intended to be maximally neutral on the nature and distribution of reasons. It can be these things for it has distinct, first-order, goals from Brandt's metaethical project. It helps us answer a very particular part (and certainly not all) of the question 'given there are reasons in such-and-such places what should we do?', not the metaethical 'what are these reason-things?' or the moral-epistemic 'and where can these reason-things be found?'

There are, perhaps, some limits to how neutral I can workably be on what reasons are or where they can be found. Most notably inserting relevance as I've defined it (or anything contingent on this definition) into one's definition of reasons can create troubling circles.

For example, the understanding of relevance offered here is ultimately intended to help make sense of the concept of brainwashing (see §6.3), by making clear why we accuse some influencing practices of 'brainwashing'. If one defines reasons or appropriately related things ('goods' perhaps, or 'what's good for X') by reference to brainwashing, in some way[222], there's a risk of circularity.

Such circularities only arise when you define reasons or appropriately related things using relevance in the sense here discussed (as, perhaps, with Brandt himself[223]) or something somehow defined in terms of relevance in the sense here discussed. Other understandings of reasons – and many are available[224] – won't generate such issues. If you accept Korsgaard's view, say, that you find reasons wherever you find impulses approved by agents' constitutive self-legislative reflection[225], then you may combine a resulting account of what things are reasons for what for whom with my position to generate claims about influences' pro-tanto legitimacy. Provided one's account of reasons renders them meaningfully presentable or representable this sort of application of my view given an understanding of the nature and distribution of reasons should always be possible.

It should also be noted that I don't reject or disesteem the influencing of behaviour by beliefs. On the contrary: as beliefs can represent reasons on this model their influence on behaviour may be valorised (in appropriate circumstances) as representations of reasons *just like any others*. The hard thing with beliefs – *as with all representations* – is only them being accurate, successfully representing justifying reasons to change one's dispositions in the ways that they seem to do. Provided they do this and owe their influence only to their doing this whatever influence they have on dispositions will meet my rationalizability requirement.

What I do reject is any claim that a person's beliefs are the *only* sorts of representations that can acceptably influence their behaviour. I instead believe *all* representations – and indeed any actual presentations – of relevant reasons can acceptably inspire changes in dispositions whatever effect they do or don't have on beliefs and irrespective of whether their content is believed or they are themselves beliefs. Inasmuch as this is the case I hold that it can be legitimate (in my limited, pro-tanto sense) to represent something to someone who doesn't believe said thing and thereby influence their dispositions, provided the thing you represent to them is a reason for them to change

---

[220] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.112

[221] D.Sobel (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492; p.476 (footnote)

[222] *see for example* C.Rosati (1996), 'Internalism and the Good for a Person', *Ethics*, vol.106, no.2, pp.293-326; pp.306-309

[223] D.Sobel (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492; p.476, p.476 (footnote)

[224] *see for examples* T.M.Scanlon (2014), *Being Realistic about Reasons*, Oxford, Oxford University Press; pp.3-14

the relevant dispositions and that it owes its effect to its being represented rather than other features of its influential expression (such as its rhetorical framing). Sometimes it's okay to inspire someone with a reason they disbelieve.

For example, imagine someone – call him Jerry – who believes that seatbelts never protect people in car crashes. He thinks the data suggesting they do is misinterpreted and all anecdotes about seatbelts protecting people are fake and he never wears one. Suppose Jerry sees someone's life saved by a seatbelt. He witnesses a crash where a passenger is thrown through the windscreen and perishes whereas the driver is clearly held safely by his belt. Jerry is a stubborn sort and refuses to update his beliefs about seatbelts given this experience. He convinces himself that the driver was saved somehow by the steering wheel, or that the passenger was just unlucky. Nonetheless, shocked by his experience, Jerry finds himself strongly compelled to wear a seatbelt; he even starts doing it from time to time. I suggest that what goes wrong for Jerry in this case is only that he fails to properly update his belief that seatbelts never protect people, perhaps also that he becomes akratic with respect to wearing seatbelts. The influence witnessing the crash had on Jerry's dispositions towards wearing seatbelts (trauma aside) was impeccable. The crash represented a justifying reason (let's say) to wear seatbelts (the consequences if one doesn't in a certain sort of crash) and this represented reason moved Jerry to start belting up, whatever his beliefs about whether he should.

(3.10) Criticisms of Cognitive Psychotherapy applied to the Rationalizability Requirement

I've argued that for an influence N to be entirely relevant to a dispositional change C N must represent or present some reason to make change C and that this representational or presentational content by itself must entirely originate the influence exerted by N. This is a specification of Brandt's position (or something like it) adapted to different ends. Given this, it risks inheriting criticisms of Brandt's position.

Velleman, notably, offers three arguments against Brandt's way of defining the good which prima facie may be adapted to challenge my rationalizability requirement[226]. Of these arguments I think the first does not challenge the position offered here (whatever challenge it poses to Brandt), the second underestimates the range of Brandt's cognitive psychotherapy, and the third – while it highlights an important nuance – requires a mistaken rejection of hierarchy amongst representations.

Velleman's first argument against Brandt's position, which would seem to adapt readily into an argument against my position, suggests that there are some noncognitive means of influencing one's motivations which allow us to achieve 'inimitable' changes in ourselves which we c*ould not otherwise achieve* (say through Brandtian cognitive psychotherapy)[227]. If any of these changes are worth making, then it follows from this that Brandt's perspective leaves something out. There are, it turns out, paths to better selves which are worth taking and which are excluded without substitutions by Brandt's model. Worse, Brandt seems to only be able to beg the question in response to this argument, insofar as for him whether a change is worth making depends ultimately

---

[225] C.Korsgaard (1996), *Sources of Normativity*, Massachusetts, Cambridge University Press, pp.112-113
[226] *see generally* J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.357-371
[227] *ibid.* p.357

on its connection to the facts, as testable through cognitive psychotherapy, and hence any change only makable through noncognitive means is presumed not-worth-making[228].

Adapting this to my position, one might argue that there are moral-dispositional changes that are worth making which may only be made by means other than those legitimated by my standard. For example it might be that I ought to practice running, but there are no relevant reasons I could be shown (however vividly) able to motivate me to practice running. Thus, it would seem I ought to make myself such that I practice running by some means other than showing myself presentations or representations of reasons to run. Inasmuch as this is the case, it turns out that my account fails to capture part of what ought to determine our decisions about influences on dispositions.

This isn't an effective criticism of my rationalizability requirement insofar as it doesn't claim to capture *all* of the things that ought to determine our decisions about influences on dispositions. It rather claims to capture only *part* of what should determine all such decisions, a defeasible consideration that applies to each of them. I can thus accept the possibility of moral-dispositional changes which should be made and which may only be made by methods which the rationalizability requirement opposes. It certainly isn't the only thing that matters, after all.

An analogous response may be made to a criticism of cognitive psychotherapy owed to Gibbard, applied to my view. Gibbard notes that Brandt must counterintuitively reject our ordinary willingness to ignore relevant reasons as a way of managing our behaviour. An honest civil servant, say, might ignore the benefits of corruption (avoid representations and presentations of them) precisely because he thinks they will corrupt him and he thinks he shouldn't be corrupted. For Brandt this ordinary way of thinking seemingly gets things wrong; the civil servant ought to attend to the benefits of corruption for only in this way may he attain 'rational' desires tested by relevant experience. For Gibbard the ordinary way of thinking should be vindicated[229]. The need to vindicate this same ordinary way of thinking can't ground criticism of my position. As I don't claim my view captures everything that ought to determine our decisions about influences – just part of what should determine such decisions – it leaves room for the kind of judgement made by Gibbard's civil servant. Generally, in circumstances where it clearly matters more that one not be led astray and less that one's dispositions be shaped by reasons, and there are reasons to go astray and one would be led astray by perceiving these reasons, one may rightly close one's eyes. I'll discuss cases of this sort in detail in Chapter 4.

Velleman's second argument against Brandt begins by noting that our existing dispositions modulate our experiences and their effects upon us. A brave person will be moved differently by a storm compared to a coward, a fickle person more readily than a steadfast one, etcetera. The effects of influences upon us will vary with our background dispositions for responding to influences. In building his account of what's good for one from those desires one is disposed to acquire through confrontation with the facts *given one's extant biases*, Velleman charges, Brandt commits to these background dispositions in a way that gives them a problematic immunity from challenge. One cannot, in Brandt's model, ask whether one should be softer-hearted (say), and hence more

[228] *ibid.* pp.357-360
[229] A.Gibbard (1990[1992]), *Wise Choices, Apt Feelings*, Cambridge MA, Harvard University Press; pp.20-21

influenced by certain facts, save by confronting said facts in one's extant hard-hearted state. This precludes meaningfully scrutinising hard-heartedness itself; it's instead *granted a strange and special protection from scrutiny in virtue of being a disposition governing one's responses to influences*. This seems to get things wrong; our dispositions for responding to influences shouldn't be specially protected in this way[230].

One analogously might argue that, on my account, certain dispositions which govern responses to influences will be similarly protected. By analogy, suppose one is a hard-hearted sort who isn't easily moved by others' suffering. What sort of reason-showing could change such a character so that they no longer have such dispositions?

I have difficulty grasping the force of Velleman's complaint here, as applied to Brandt's cognitive psychotherapy or my position. Cognitive psychotherapy does not, as I understand it, grant special protection to dispositions for responding to influences. It takes people as it finds them (as Brandt insisted[231], so as to avoid the detachment from people he attributed to ideal observer theories[232]), sure, but Velleman is wrong in thinking that it renders extant dispositions for responding to influences specially and troublingly beyond challenge.

For example, Velleman considers a hard-hearted sort, unmoved by the plight of others. Velleman writes 'in asking whether it would be good to be softer of heart, we needn't be asking whether the facts would soften us, given how hard-hearted we are'[233]. Consider what sort of relevant facts a cognitive psychotherapy program might expose one to to test one's hard-heartedness. Not all these facts need be such that one's response to them will be hostage to one's extant hard-heartedness. One set of relevant facts one might be exposed to, say, may show how hard-hearted people perform socially – their tendencies to alienate people, climb certain hierarchies, and so on. One's extant hard-heartedness needn't modulate one's responses to such facts.

Granted, some things relevant to one's hard-heartedness may be such that one will be more or less moved by them on account of one's hard heartedness. Such modulation, though, is a ubiquitous feature of influencing rather than the result of some special and problematically protected class of extant dispositions inherent in Brandt's – or my – account. Whenever one is influenced by almost anything in almost any circumstance how one responds to said influence will be partly the result of one's extant dispositions. For example, if one has somewhat durable existing dispositions regarding certain behaviour which must be dislodged by an influence in addition to inculcating new dispositions regarding said behaviour one might expect this influence to have more difficulty achieving results. Such durable existing dispositions may exist in practically anybody governing practically any behaviour. Some of us are 'set in our ways' about some things. If the possibility of extant dispositions modulating the effect of influences is ubiquitous, though, it isn't clear that there exists any special class of dispositions for responding to influences specially protected from challenge, as Velleman worries[234]. *All* our dispositions may be somewhat protected from challenge

---

[230] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; p.360
[231] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.113
[232] *ibid.* p.225
[233] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; p.360
[234] *ibid.* pp.360-362

based on our individual peculiarities. All influences on dispositions risk encountering resistance resulting from existing dispositions, not just influences on dispositions for responding to influences.

While it may undermine Velleman's second argument against Brandt this risk seemingly strengthens his third, centred on what Velleman calls the 'Problem of Representation'.

Velleman argues that there exists indeterminacy in connections between represented relevant facts and behaviour. Not only are our responses to representations of relevant facts modulated by our extant dispositions, and thus different person-to-person and time-to-time, but these responses differ from representation-to-representation. For example, an open-heart surgery might be represented to one through mental imagery, flow chart or narration, each time differently moving one. The same surgery might also be represented to one formally or informally, micro- or macroscopically, and so on, each time differently impacting one's response. This kind of impact, ubiquitous to the business of representation, means that there's no determinate response connected with any fact for the medium of a fact's representation is too determinative of peoples' responses to it. This being the case, Velleman argues (following Railton[235]), Brandt's use of cognitive psychotherapy to ground claims about what's good for one (at least how it may be discerned, perhaps what it is[236]) renders the good either empirically undeterminable or even downright indeterminate, neither of which are acceptable[237]. Furthermore, there's no hierarchy of representations available to help by declaring some representations more representative than others, for any such hierarchy would have to differentiate between distinct compatible representations, a feat impossible unless (implausibly) reality is somehow biased in favour of representational styles ('pictorial or verbal, technical or slangy, close up or far off' and so on[238])[239].

This argument may be adapted to the account I've offered insofar as my account valorises influences on dispositions which represent relevant facts without detailing how the relevant facts should be represented. Given this, one might argue that my account is vulnerable to Velleman's concerns here. Even if one represents relevant facts to somebody to influence them, if the medium of influencing more determines the effects of this influencing than any represented facts (reasons) why should we valorise these facts' role in this influencing? If the alienation of one's dispositions from relevant reasons is just an inescapable result of being moved by representations, not something one can combat by prioritising the influence of certain representations[240], then why should we even try to combat it (and how could we even do so?)?

There is, however, some hierarchy of representations sufficient to justify claims that some representations are better than others as influences (though there may be no such hierarchy prior to reasons and hence no such hierarchy sufficient for Brandt's purposes, distinct from mine). Mediums aside, clearly some representations are more representative, at least with respect to those things

---

[235] *see* P.Railton (1986), 'Facts and Values', *Philosophical Topics*, vol.14, no.2, pp.5-31; pp.19-25
[236] D.Sobel (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492; p.476, p.476 (footnote)
[237] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.368-369 (footnote)
[238] *ibid.* p.366
[239] *ibid.* pp.365-366
[240] *ibid.* p.366

that ought to impact specific dispositions. To use Velleman's example[241], much more of what it is to undergo open-heart surgery, relevant to one's dispositions towards such surgery, will be represented in a relevant chapter of a good medical textbook than a photograph of such surgery taking place. It might not be all the same stuff – one might see things in the photo one couldn't in the textbook – but, I suggest, it will plausibly be more of *what one ought to consider in deciding whether one ought to have such a surgery*. Given this, it's fair to describe the representation offered by the good medical textbook as better (in the sense that it represents more of the field of reasons one should consider in deciding whether to have open heart surgery) than the representation offered by the photo as an influence on dispositions with respect to one's willingness to undergo open heart surgery. Generalising, it's possible for representations to represent more or less of the relevant contents of reality (more or less relevant reasons) and in doing so become better or worse influences on dispositions. As I argued in §3.7 representing more relevant realities is always a better-making feature of influences. This being the case there does exist some hierarchy of representations sufficient to make some influences better than others.

I suggest it's this sort of hierarchy Brandt was highlighting in invoking the concept of 'vividness[242]' (which I use for similar purposes); a medium-independent hierarchy of relevant detail. Contra Velleman[243] this hierarchy needn't have at its top (or dominating its upper echelons) some implausible optimal medium; for all we know it may have multiple media able to maximally represent relevant realities to agents able to successfully interpret said media (a photo isn't going to represent much to the blind man in the presently significant sense, say, or at least as much as the same photo's audio description).

This may or may not defend cognitive psychotherapy from Velleman's third argument. With respect to my separate project, however, I think this counterargument is sufficient.

In defining relevance in §3.9 I noted that the influence of non-representational parts of influences can modulate their effects and that this imperils their legitimacy on my account. Influences ought to influence by showing relevant reasons, not through whatever irrelevant or non-representational content adulterates these showings. This part of my position helps me cope with Velleman's 'Problem of Representation'. The more an influence in depends on features it has other than its representation of relevant reasons (such as its non-representational aesthetic qualities) the less legitimate the influence is. The less legitimate the influence is the weaker the pro-tanto reason we have to tolerate its effects on dispositions until, eventually, we gain pro-tanto reason to prevent its effects in the case where it owes most of said effects to features which do not represent relevant reasons. By this standard my account may accommodate Velleman's concerns about the modulation of the effects of representations by non-representational (and indeed irrelevant) features.

This accommodation extends to the extant peculiarities and dispositions of influence-subjects and these peculiarities' modulation of influences. Extant peculiarities and dispositions like influence-subjects being inattentive, in good moods or 'set in their ways'. Insofar as such things – which are

---

[241] *ibid.* pp.365-366
[242] R.Brandt (1979), *A Theory of the Good and the Right*, New York, Oxford University Press; p.111
[243] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.367-368, 370

certainly not representations of relevant reasons by influences – may modulate the effects of influences on dispositions on my account they may render influences less legitimate in proportion to this modulation. It follows from this that appropriately peculiar influence-subjects may, on account of their peculiarities, be such that some otherwise legitimate influences may only illegitimately influence their dispositions (I'll discuss such examples in detail in §4.5).

Velleman could attempt to resist my counterargument (offered in the preceding five paragraphs) by leveraging his claim that facts-in-themselves or perceptions of them have 'no single motivational impacts' associated with them[244]. My counterargument, after all, trades on the idea that some representations are more representative (vivid) than others – in a way sufficient to ground a hierarchy amongst influences – for the reason that they more closely approximate the motivational impacts of the relevant facts (either by representing more of them or representing them with less non-representational or irrelevant adulteration). This standard is premised on certain facts-in-themselves, certain reasons, having associated motivational impacts (i.e. on these facts, when perceived, tending to impact people in certain ways). If this is strongly[245] denied it isn't clear how this counterargument could work.

How, though, can one prove that facts-in-themselves do not have single (or – sufficient for my counterargument – sets of) associated motivational impacts? We humans may or may not be able to perceive facts-in-themselves and we certainly very often deal with only representations of them. It's also true that these representations can have diverse motivational impacts. One cannot, however, infer from diversity in the motivational impacts of representations of facts that facts-in-themselves lack single associated motivational impacts. This diversity may be explained by diversity in representations or diversity in influenced parties. To go beyond this and claim that facts lack single associated motivational impacts (as Velleman seems to[246]) is to infer beyond evidence. For all we know the facts themselves might be quite consistent in what they do to dispositions; it's only us and our ways of representing them that diversify our responses (in ways which, as argued in preceding paragraphs, do not defeat my view).

Indeed there seems to be some evidence that facts themselves can have single associated motivational impacts (or sets of impacts), for different representations of the same fact *tend* to cause similar motivational impacts.

Imagine an image of Elizabeth the First torturing someone. The changes in your dispositions motivated by this image might depend on many variables – your background opinions of Elizabeth the First, the credence you give to the image, the mood you're in, etcetera. Amongst these things will be the representative content of the image; what it shows you. This content might, say, make it the case that on perceiving said image one becomes more negatively disposed towards Elizabeth the First. One becomes more likely to describe her as cruel, say. This effect of content upon impact is a general feature of representations and seems to be constant across different representations which

---

[244] *ibid.* p.366

[245] A weak denial won't work against my counterargument; if facts associate with motivational impacts *a bit* then we may still make claims about how representative representations are based on how well their motivational impacts track the motivational impacts facts associate a bit with.

[246] J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; p.370

share representative content. If a verbal description of Queen Elizabeth the First torturing someone was supplied to one, representing the same content as the putative image (with neither additions nor subtractions, as an accurate caption might), we should expect it to impact one in the same way as the image. This impact might not have the same exact extent and might be differently modulated by one's peculiarities (one might be more moved by images than descriptions, say) relative to the impact generated by the image, but we would expect it to be *similar*, to move one 'in the same direction' on some behavioural axis. We might expect one to still become more negatively disposed towards Elizabeth the First, say. We'd find it strange if someone was motivated by the picture to dislike Elizabeth and motivated by the description to like Elizabeth, if they both showed the same representational content.

This is all the case whether the things represented are true or false. Whether the putative picture of Elizabeth and the description represent the same real events or the same fiction we will expect them to have motivational effects which are similar (at least on agents not biased somehow against either representation). This is the case because the content of a representation is somewhat determinative of its impact irrespective of its truth (though perhaps not opinions of its truth).

This could not be the case were there 'no single motivational impact associated with the facts' for amongst the contents captured by representations are sometimes facts (plausibly, some representations are true). If contents somewhat determine the impacts of representations and some contents are facts it follows that sometimes facts determine the impacts of representations. Were 'no single motivational impact associated with the facts' then in all such cases we should expect to not see the kind of tendency for impacts to converge that I've described. Such convergence would imply a kind of connection between facts and motivational impacts inconsistent with Velleman's potential rejoinder. We should instead expect the truth of representations of a fact to, spookily, make it the case that the motivational effects of these representations diverge further than one might expect given their shared content. This clearly doesn't happen; motivational effects rather seem to depend upon – and converge somewhat if they share – content, *irrespective* of this content's factuality. This being the case there seems to be space for facts to associate with motivational impacts. If facts can do this, though, Velleman cannot deny such association to defend his 'Problem of Representation'.

Velleman could respond to this argument by suggesting fictions, as well as facts, don't associate with motivational impacts. In responding in this way, though, Velleman would have to commit to the view that contents generally (factual or fictive) do not associate with motivational impacts. In claiming this Velleman would have to deny that shared representational contents lead to any convergence in responses to representations, which would be a very odd thing to deny. Some such convergence seems to be what we would expect in my tortuous Elizabeth example. We'd also expect it in the case where, say, an event is described to one (a bilingual) in a given way in English or counterfactually in the same way in French (as it might be by an accurate translation). It would be odd were all such and similar expectations wrong, as they would have to be for contents not to associate with motivational impacts generally (at least in a strong enough way to undermine my counterargument).

Thus Velleman's 'Problem of Representation' is no problem for me. It's true that representations may be modulated by their non-representational content or the peculiarities of their percipients but

neither of these possibilities threaten my position so much as constitute the kind of corruption of influencing that it opposes. It's also true that my account depends on an understanding of the vividness of relevant representations which demands that representations of facts have associated motivational impacts but they do seem to have such impacts, as with contents of representations generally. Furthermore my account doesn't give dispositions for responding to influences any implausible unique protection against being influenced. Nor does it claim to capture all the ethics that might impact judgements about influencing such as to preclude, with sufficient justification, the use of illegitimate methods to achieve results that can't be got legitimately. This all being the case, I take it that the position I've developed through engagement with Brandt's model is safe against the analogization of at least some criticisms of this model to my position.

(3.11) Rationalizability as a Value in the Ethics of Changing Moral Dispositions

I've argued that the relevance relations that can exist between changes in dispositions and influences are best understood rationalistically. More precisely, I suggest that for an influence N to be relevant to a change in some disposition(s) C, N must represent or present some justifying reason to make change C. Further, the more that this representational or presentational content is responsible for C the more relevant N is to C in this sense.

I've also argued that we value such 'relevance' in making decisions about the acceptability of changes in at least our moral dispositions. In §3.7 I explained this value with reference to a general need for our moral behaviours to be causally responsive to the world we inhabit, explained though not justified by the historic evolutionary advantages of such dependence. This need compels us to prefer inspiration from the right bits of the world: the 'relevant' ones. To find inspiration – influence over behaviour – elsewhere, in 'irrelevant' things and representations of such things is to neglect this need and become alienated from what ought to inspire one.

Putting these points together says something about how to go through life, something about the ethics of changing moral dispositions. Put simply, on this model, absent sufficient reason to do otherwise one should work to limit moral-dispositional change to those changes which causally respond to presented or represented reasons. It is by doing this that we can avoid behaviour becoming alienated from those things that ought to determine it. It is by doing this that we can ensure that processes of moral-dispositional change (whatever they may be) do not ignore those things of moral importance that ought to lend them whatever power they acquire over moral dispositions. Rational or otherwise, it is my view that moral-dispositional changes ought to be 'rationalizable': rightly explained in terms of the effects of perceived reasons on dispositions. The 'reality requirement', earlier suggested, subsumes into this 'rationalizability requirement'.

I grant that in practice representing or presenting some relevant reason to make a given moral-dispositional change – and hence influencing in the way I say one should – needn't be easy to achieve or confirm. For one thing the reasons which must be represented or presented by (pro-tanto) legitimate influences on this model (like reasons generally) are only ever reasons given contexts which make them reasons. These reasons might also be plain obscure, with complex and hard-to-show characters. In such situations representing or presenting such reasons successfully may be thought – in practice – to require onerous detail, contextual and otherwise. This presents us with a question: how much of what makes a reason a reason – *how much of a reason*, really (or, one might say, the state of affairs that makes the reason a reason) – must an influence represent to represent the reason and attain rationalizability?

Something must be said here for if the answer to this question is too demanding – if, say, rationalizable influences must represent reasons *fully* including the entire context that makes them reasons – then the rationalizability requirement I suggest may make strange and onerous demands. It may demand we prefer influence by representations or presentations that are beyond us and our capacity to show and perceive reasons. This would give us reason to reject the requirement.

Recall Annie. I claimed in discussing her that a news report showing 'real people suffering real hardship and in need of help' would be a relevant and (pro-tanto) legitimate influence on Annies' dispositions towards giving to charity. What must a news report detail to accomplish this representative task? This depends on the character of the reasons the representation must portray. It might, say, require that the news report details the character of the suffering involved (are people hungry, thirsty, or overworked? To what extent?). It might require some illustration of how Annie's increased charitability might help said suffering people or those who are like them[247] (might it help feed them, supply water, or lobby for better conditions?). It might require some contextual nuancing of what the report shows (are the people hungry or thirsty because their land is overpopulated, is their overwork benefitting them?). It might, in some circumstances, merely require the utterance of the true sentence 'people are suffering hardship and need help'. What matters in any case is that the report represents enough of the right parts of the facts that somebody who only takes proper reasons (whatever they are and wherever they are) into account in their judgements, were they to believe the content of the report, would be able to more-or-less properly factor this content into a judgement about making the dispositional changes perceiving the reports can cause. *What matters is that the reports clearly detail enough of the right parts of the facts that they show whatever reason for moral-dispositional change these facts constitute, such that it may act upon the percipient with roughly appropriate force* (whatever this force is).

In some circumstances this might require significant detail or even detail expressed in certain ways which may be determined by the peculiarities of the influenced party (English words for English monoglots, say). There may be cases where it may be achieved by a single stated sentence (where, say, the influenced party is able to derive a rich meaning from said sentence for some reason). There are certainly also cases where it isn't achieved at all. Importantly – such complexities aside – we don't generally think that representing or presenting reasons in sufficient detail, with sufficient context, and in the right ways to allow these representations or presentations to capture reasons' appropriate force is impossible or impractical. We might debate the separate question of what the appropriate force of a given reason is but – obscure reasons and adverse circumstances aside – we don't regard reasons as impossible or impractical to show to people. At least, we seem to regard our armoury of senses and symbols as sufficient to allow us to present or represent reasons (and interpret these presentations and representations) with enough reliability to use them frequently in, say, talking to each other (whether we mean this talk to alter ideals, influence dispositions, or both).

This being the case, the rationalizability requirement shouldn't often generate onerous demands which outstrip what's representationally or presentationally feasible.

It also seems, then, possible for us to determine whether influences show reasons successfully enough for them to satisfy the rationalizability requirement. Case-by-case, we can be workably sure whether an influence shows a reason to somebody, depending on what it says in a manner said somebody can grasp. Contingent on this, we can also case-by-case be workably sure whether

[247] For whom the particular individuals shown to Annie may act as representative metaphors, in the way I suggested Dickens' characters might in §3.2.

influences alienate from relevant reasons, and are thus rendered illegitimate (and such that we have pro-tanto reason to oppose them) given the rationalizability requirement.

Granted, it's not always perfectly clear whether a given method of moral-dispositional change alienates from relevant reasons or not. Drugs offer a good example of this problem. An 'altruism drug', were an effective one to be made, would have to work by somehow 'opening the eyes' of users to features of reality relevant to altruism for its administration not to constitute a blameworthy alienation of the subject from the reasons that ought to move them. It's not clear that existing 'altruism drugs' work in this way.

For example, some psychologists say that the oft-suggested 'altruism drug' oxytocin (a hormone which can be pharmacologically boosted) works by enhancing one's ability to empathise[248], while others say its main action is to enhance interpersonal bonding feelings[249]. This debate is, given the rationalizability requirement, important to assessing the acceptability of oxytocin as a tool for modifying moral dispositions.

If the drug is an empathy enhancer it plausibly works by bringing subjects into presentational or representational contact with feelings of others plausibly relevant to altruistic behaviour, for empathy is a skill for knowing (through some sort of mental presentation or representation) others' feelings – things plausibly relevant to one's altruism (at least in many circumstances)[250]. As such if oxytocin is an empathy-enhancer then in virtue of the ethics of moral-dispositional change it seems to be an acceptable tool for manipulating moral dispositions or rather an acceptable enabler of such manipulation (of a sort I'll discuss in §4.5). It is, on this view, no more problematic than giving someone glasses that allow them to better see the reasons that ought to motivate change in their moral dispositions.

On the other hand if oxytocin is a bonding-enhancer then its ethics are more precarious. Proponents of the bonding analysis claim that oxytocin enhances altruism only towards those one already feels bonded with, like members of one's family, tribe or race. It can be associated with discrimination against 'outsiders[251]'. Plausibly 'being an insider' isn't relevant to whether one is a fitting recipient of altruism in the same way as one's feelings are relevant (I'll discuss alternative possibilities, and how they relate to the rationalizability requirement, in §6.5). Thus, if oxytocin is a bonding enhancer, then on the proposed model it pro tanto isn't an acceptable tool for achieving moral-dispositional change. Oxytocin isn't such an acceptable tool insofar as it doesn't work by presenting or representing things relevant to the altruistic dispositions it engenders (or enabling such presentation or representation).

It's an empirical matter whether oxytocin works by enhancing empathic perception or by deepening feelings of bondedness, and I can't settle such questions here (for all I know it might do both). What I hope I've shown, though, is how the account of the ethics of changing moral dispositions that I've

---

[248] N.Levy, T.Douglas, G.Kahane, S.Terbeck, P.J.Cowen, M.Hewstone and J.Savulescu (2014), 'Are You Morally Modified? The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-171; p.117

[249] *ibid*. p.118

[250] J.Zaki, J.Weber, N.Bolger, K.Ochsner, and M.I.Posner (2009), 'The Neural Bases of Empathic Accuracy', *Proceedings of the National Academy of Sciences of the United States of America*, vol.106, no.27, pp.11382-11387; p.11382

[251] *see* C.K.W.De Dreu, L.L.Greer, G.Van Kleef, S.Shalvi and M.J.J.Handgraaf (2011), 'Oxytocin Promotes Human Ethnocentrism', *Proceedings of the National Accademy of Sciences of the United States of America*, vol.108, no.4, pp.1262-1266

arrived at can help us address the normative component of such debates. It can tell us something about what we should do about – and with – influences such as oxytocin or this-or-that given advert or speech, given an understanding of their workings.

There are, however, many more considerations that must be factored into judgements than those considerations detailed in this chapter. Having ones 'eyes opened' by an empathy drug might pose no problem with respect to the ethics of moral-dispositional change, but we might still worry about the sort of relationship this creates between us and the drug supplier. We might also worry that under such circumstances we might lose or surrender something of our self-control. I'll address such worries (which permeate extant discussion of the ethics of influencing), most often articulated with reference to some sort of *freedom*, in Chapter 5, building towards broader conclusions in Chapter 6.

For now, I can only offer some preliminary indicative comments about the ethical program implied by the rationalizability requirement.

My account (among other things) identifies a certain pro tanto reason, general to moral-dispositional influencing cases, to reject moral-dispositional influencing (on oneself or others) by things that neither represent nor present reasons to make the changes in question. In contemporary society this view alone licenses some changes in how we influence and engage with influences.

President Kennedy said 'in a free society art is not a weapon … Artists are not engineers of the soul', before darkly adding 'It may be different elsewhere[252]'. Given the advancement in means of influencing I highlighted in §1.1, plausibly we're presently moving to something more like Kennedy's 'elsewhere' where 'soul engineering' is ordinary business, not least for artists. One need only watch an average television commercial to be bombarded by irrelevantly beautiful people encouraging the consumption of various goods to the tune of irrelevant music while flaunting lifestyles and ideals irrelevant to said consumption. It is almost cliché to point out that modern commerce doesn't sell goods with description, it sells with inspiration; exactly the sort of inspiration I argue shouldn't be traded on the market by those who can make and manage it. It should instead – if I'm right – be left where it's found, attending only those 'relevant' parts of reality that deserve it. Leave the attractiveness of actors for attracting mates or making beautiful art; don't use it to sell cars. Resist – by appropriate means – anyone using it to, say, sell you a car or sway your vote or conscience[253].

These are, though, only the issues of the present. The more-or-less immediate future seems to hold further challenges. Psychologists have long institutionalised cataloguing the many levers that can be used to manipulate human behaviour; admen have only availed of some of them so far. Some, famously B.F. Skinner in a previous generation[254] and Cass Sunstein more recently[255] have argued that the widespread use of these levers by enlightened leaders could do good. Those who argue in favour of 'biomedical moral enhancement', such as Persson and Savulescu[256], undertake the same intellectual project citing variant means. For all versions of this project the rationalizability requirement presents a challenge; perhaps a defeasible challenge, but one which demands answer. I'll sketch some of the contours of this challenge in Chapter 6.

---

[252] J.F.Kennedy (1963), 'Remarks at Amherst College', Event in Honour of Robert Frost, Amherst, Massachusetts, October 26 1963

[253] One may regard the dispositions involved in some of the cases I mention here as non-moral and thus beyond the remit of the position I've defended. In §6.4 I'll argue this is less often true than one might imagine.

[254] B.F.Skinner (1971), *Beyond Freedom and Dignity*, London, Penguin Books; pp.9-11

[255] J.J.Chriss (2014), 'Influence, Nudging and Beyond', *Society*, vol.53, no.1, pp.89-96; pp.89-90

[256] *see* I.Persson and J.Savulescu (2012), *Unfit for the Future*, Oxford, Oxford University Press

Before this may be attempted, however, it's necessary that I detail how the standard captured in the rationalizability requirement may be operationalised. How should we minimise the alienation of our moral-dispositional change from the reasons that ought to inspire it, given that other things matter and we're not always agreed about what these reasons are? Furthermore, how does this minimisation relate to protection of freedom often thought important in placing restrictions on influencing? I'll try to answer these questions in Chapters 4 and 5 respectively.

(3.12) Conclusion

The rationalizability of an influence on any number of moral dispositions is always a better-making feature of the influence, significant to its ethical evaluation. This rationalizability requires that influences be relevant to the dispositions they influence. This relevance requires, for any influence N that for it to be relevant to a change in some disposition C, N must represent or present some reason to make change C. The more this representational or presentational content of N is responsible for whatever causal impact N has on C the more relevant N is to C and the more rationalizable we may say N insofar as it effects C. Provided N is rationalizable at all in this way it's effect on C gains a certain pro tanto legitimacy; the more N is rationalizable in this way the more of this legitimacy this effect acquires, the more acceptable said effect on C becomes (ceteris paribus). Where this relevance is lacking influences lack this same legitimacy and must be justified or else rejected. This set of standards, composing a 'rationalizability requirement' on moral-dispositional influencing, advance understanding of the ethics distinctive to evaluating moral-dispositional changes (whether these changes be orchestrated by the changed party or others). They resist criticisms of analogous standards owed to Richard Brandt suggested by Velleman and Gibbard. In conjunction with sufficient knowledge of how influences work they may be used to evaluate influences on moral dispositions in a novel and subtle way.

Chapter 4: Towards Application

(4.1) Abstract

In this chapter I show how my rationalizability requirement may be applied to real-world problems characterised by unknowns and moral complexity. I show how the requirement can still yield insights despite disagreement about the reasons it depends upon and the presence of competing values in practical contexts. No matter how reasons are generally distributed, some things clearly are or are not reasons to adopt certain dispositions and certain influences lack presentational or representational content. Thus, I suggest, we may find some agreement about which influences are legitimate despite disagreement about reasons. Where dilemmas between the demands of the rationalizability requirement and other moral concerns occur, I argue we should weigh these demands in proportion to how many agents are being influenced, how effective the methods of influencing involved are, and how strong the reasons behind these demands are. I argue, however, that true dilemmas between the demands of the rationalizability requirement and other moral concerns only occur in specific circumstances, as moral reasons that threaten to override the rationalizability requirement themselves enrich the influencer's capacity to influence legitimately.

(4.2) Introduction

I've made claims about how we should draw a line between legitimate and illegitimate influences on our moral dispositions, such that we have pro tanto reason to accept the former and oppose the latter. I suggest that we find some things, things which show reasons to make certain moral-dispositional changes, relevant to these moral-dispositional changes. Dispositions ought to be influenced by these 'relevant' things; they ought not to be influenced by irrelevant things. Acceptably violating this rationalizability requirement by suppressing relevant influences or accepting (through toleration or use) irrelevant ones requires justification.

Applying this account generates two related fundamental problems.

The first problem stems from the account's reliance on reasons. People disagree about reasons. Some say some things are reasons for certain dispositional changes, some say the same things are not. Some say that some things are reasons for some but not others, that reasons are relativistic in character[257], while others claim for them universality[258]. Such disagreements threaten attempts to derive insight from my account. If we can't agree what's a reason for what for whom, how may we helpfully assess the legitimacy of influences by asking whether they constitute reasons? Surely any such effort is doomed to entirely reduplicate debates about reasons? If this is the case, it isn't clear how much help the rationalizability requirement could ever be.

Presuming that this problem can be managed my account must deal with a second. The rationalizability requirement doesn't operate in a moral vacuum; per moderate instrumentalism its demands must be factored into judgement alongside others, as with any partial account of what's valuable. How, however, can this 'factoring in' take place? How important is influencing moral dispositions in the right way, in a world in which life, limb and livelihood can depend upon people having the right moral dispositions?

---

[257] B.Steinbock (1981), 'Moral Reasons and Relativism', *The Journal of Value Inquiry*, vol.15, no.2, pp.157-168; pp.158, p.162

[258] *ibid.* pp.165-167

These two problems must be addressed before my ideas can be used to critique practice.

I think the first of these problems may be mitigated by highlighting how my account creates space for agreement about which influences are legitimate even given disagreement or uncertainty about what's a reason for what for whom. As to the second problem, I believe that the normative force of the demands of the rationalizability requirement is proportioned to the count of those who are subjected to influences at issue, the effectiveness of these influences, and the importance of dispositions being changed. These criteria, I think, should help us resolve dilemmas between the demands of the rationalizability requirement and other concerns. I also suggest that due to reasons to override the rationalizability requirement tending to make it easier to meet it (by giving one the sort of reasons for dispositional change one must show someone in order to legitimately change their dispositions) many clashes between the rationalizability requirement and other concerns will be more soluble than they first appear.

(4.3) Reasons and the Rationalizability Requirement

Reasons play a foundational role in my positive thesis. I argue that what makes the difference between legitimate and illegitimate influences on behaviour (in the limited sense I discuss) is whether they present or represent reasons to change behaviours in question and whether this content does the influential work. As such my account is only useful alongside some prior list of reasons. This creates problems.

Such lists can invite, at least, questions about what's a reason for what. One person might find the suffering of an underclass sufficient reason to change his voting behaviour to support tax-and-spend fiscal policy, another might find this suffering no reason at all for such change[259]. These sorts of disagreements may be results of misunderstandings or ignorance; one of the characters in such a case may be mistaken about the realities of the situation in question. When this is the case the question of what's a reason for what (sufficient or otherwise) may be resolved by debate and reasoning about the problem in question.

More vexing though, prima facie, is when these methods fail and the disagreement becomes 'fundamental', seemingly grounded in divergent fundamental values rather than mere ignorance[260]. With such divergent values in play some thinkers are encouraged to adopt relativist[261] or subjectivist[262] views of reasons, whereby they diverge significantly from one person to the next, based on features of the people in question (such as their psychologies[263] or cultural backgrounds[264]).

The relativist about reasons believes that agents have reasons relative to something[265]. This is a view which they share with most universalists. The universalist might, for example, say that her god tells everybody not to steal, and this gives them a reason not to steal. If this is the case, though, the reason that we have not to steal remains relative to something: divine will (in this example). Were this will different, the reason in question would be different or non-existent. This trivial kind of relativity is a widely acknowledged feature of reasons even within universalist traditions; the idea

---

[259] M.Michelmore (2012), *Tax and Spend*, Philadelphia, University of Philadelphia Press; pp.132-133
[260] R.Rowland (2017), 'The Significance of Fundamental Moral Disagreement', *Noûs*, vol.51, no.4, pp.802-831; p.802
[261] *ibid*. p.817
[262] P.Foot (2002), *Moral Dilemmas and Other Topics in Philosophy*, Oxford, Oxford University Press; p.190
[263] *ibid*. p.190
[264] S.Sikka (2012), 'Moral Relativism and the Concept of Culture', *Theoria*, vol.15, no.1, pp50-59; p.50
[265] M.Baghramian (2004), *Relativism*, London, Routlege; p.207

that what reasons we have for action depends on things is not an idea the relativist has any monopoly on. What is distinctive about the relativist is what they think reasons are relative to. Traditionally the relativist holds that reasons are relative to moderately shared things like cultures which the universalist traditionally denies make significant differences to the reasons agents have for action[266]. The subjectivist about reasons, meanwhile, sees one's reasons as relative to particulars of the subject which both relativists and universalists exclude from the set of determinants of reasons[267]. Universalists think that agents have reasons relative to things which are universal (like laws of practical logic or universal human nature) and thus that all agents have the same reasons for action in appropriately similar situations. Everybody thus agrees that different people have different reasons; people disagree about what the proper determinants of these reasons are and how these determinants are distributed.

The problem posed by disagreement about reasons for my account thus decomposes into two separate problems.

The first problem is saying what counts as a reason for what for whom. If we disagree about what things determine what is a reason for what for whom and we are thus not workably sure where the reasons are, how can we employ an account of influences' legitimacy dependant on knowing where the reasons are?

The second problem is the problem of whose reasons count in assessing the legitimacy of influences on an agent's behaviour. Should we look to the reasons that the influencer has, the influenced party has, or the reasons of some third party? These might all be different, after all.

It's beyond the scope of this thesis to completely answer the first of these problems, as this would require a full articulation of what's a reason for what for whom and thus a comprehensive ethics. Nonetheless I'll argue that absent a complete answer to this problem there's still a great deal we can say. As to the second of these problems it follows from the justification of the rationalizability requirement that only those reasons an agent has – whatever they are – may exert legitimate influence over their behaviour. This is not the same thing as saying that we may only be legitimately influenced by reasons we know that we have (which would be a transparency condition comparable to Cowley's position[268] that I dismissed in §3.4).

(4.4) Reason for Whom?

I'll address the second problem first.

I claim that for any influence N to be relevant to a change in some disposition(s) C, and thus a legitimate (in my limited pro-tanto sense) influence upon disposition(s) C, N must represent or present some reason to make change C. I've justified this claim by arguing that only when influence proceeds in this sort of way does it preserve a proper connection between our moral dispositions and the parts of reality that ought to shape them, the parts we call reasons. Given this justification these reasons, in turn, must in some way apply to *us*, be relevant to *our* dispositions, if they are indeed to be such that they should shape *our* dispositions. Whether they might be reasons for other people to make similar dispositional changes is irrelevant (save perhaps as a source of evidence by way of analogy, where others are like one in those ways that make reasons apply to agents). The

---

[266] *ibid.* p.209
[267] *ibid.* pp.209-213
[268] C.Cowley (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290; pp.277-278

relation that must be preserved is between one's own moral dispositions and those reasons that ought to inspire change in *them*.

Generally, for an influence to influence legitimately (in my limited sense), it must work by showing the influenced agent things that make it the case that they ought to make the moral-dispositional changes the influence causes; reasons which apply to the agent and which are relevant to *their* moral dispositions. This needn't involve showing the agent reasons they know they have (it may even involve showing the agent reasons they believe they don't have but really do; think of Jerry form §3.9).

For example, suppose that the appearance of drunkenness is a reason to get drunk less around here, but not in Boozia. Suppose somehow all Boozians ought to find the appearance of drunkenness a reason to get drunk more and not less, given their culture or what makes Boozians happy or some other determinant of what's a reason for whom. Were a Boozian shown images of someone being drunk, and thereby inspired to get drunk less this wouldn't, in such a case, be an instance of legitimate influencing. Were I – being from around here – shown the same influences and similarly inspired, this would be an instance of legitimate influencing. This would be the case because (we're supposing) my reasons entirely differ from the Boozian's and it is only influenced individuals' reasons which count (as anything other than, in appropriate circumstances[269], evidence) in assessing influences' rationalizability.

(4.5) Agreeing and Disagreeing about Reasons, Agreeing about Influences

Most perspectives on the distribution of reasons agree that there are facts or fact-like things[270] about what's a reason for what for whom (error theories about normativity notwithstanding; and no *ethics* of moral-dispositional change will ever be consistent with an error theory about normativity[271]), though these perspectives dispute the particulars of these facts and to what extent they can be generalised. It is from such disputes that §4.3's first problem with applying my account to judgement comes. If we cannot agree upon what constitutes a reason for what for whom, how can we helpfully cite facts about such matters in answer to questions about the legitimacy of influences?

As a first point, in answer to this worry, it should be realised that even if in some circumstances we can't agree what's a reason for what for whom, this needn't mean the rationalizability requirement is useless. It should still help us more generally to judge which influences are acceptable. It should do this by telling us something about what we want from influences, something which can transform unfamiliar problems into more familiar and, we may hope, solvable ones.

You and I might agree, say, that the beauty of a landscape constitutes a reason to preserve it, yet disagree about whether showing people said landscape is an acceptable way to encourage them to help its preservation. One of us might worry, say, that showing off the beauty of the landscape might be somehow unfair or manipulative. At the same time we might agree about the instrumental features of the matter; we might agree how costly it would be to show off the landscape, say, and how important it is that the landscape be preserved. If the rationalizability requirement holds,

---

[269] That is, when the determinants of reasons are shared such that what another has as reasons may evidence what influenced individuals have as reasons.

[270] G.Bex-Priestly (2018), 'Error and the Limits of Quasi-realism', *Ethical Theory and Moral Practice*, vol.21, no.5, pp.1051-1063; pp.1052-1053

[271] M.Lutz (2014), 'The 'Now What' Problem for Error Theory', *Philosophical Studies*, vol.171, no.2, pp.351-371; pp.351-352

there's something problematic about such disagreement. If the landscape's beauty is a reason to preserve it then showing off said beauty should be, other considerations aside, an acceptable way to influence people to preserve the landscape. If we agree about the former thing we should agree about the latter thing and if not we must justify failing to do so.

Generally, given the rationalizability requirement and ceteris paribus, agreement and sureness about reasons should lead to agreement and sureness about influences' legitimacy. This is useful for we have means, more-or-less effective, to achieve agreement and sureness about reasons. We might work through our understandings of relevant facts and purge them of false beliefs, or we might do practical reasoning about what facts recommend what dispositional changes. Generally, we might apply methods for improving the accuracy of beliefs about which justifying reasons agents have, and thereby attain more agreement and sureness in such matters. In doing this we might come to agreement and sureness about what reasons agents have and thus, given the rationalizability requirement, what must be shown to them to legitimately influence at least their moral dispositions. Such an inferential pathway wouldn't be obviously available absent the rationalizability requirement.

Moreover – given the rationalizability requirement – although we may not agree what makes something a reason, or what the best account of the distribution of reasons is (or even what the best way to make our beliefs about reasons more accurate is), we may still achieve agreement about the legitimacy of influences for many circumstances.

There are two subsets of these circumstances. The first of these subsets involves circumstances where influences represent or present things which are obviously reasons or not reasons. The second of these subsets involves circumstances where the effective content of influences fails to present or represent anything to the influenced party.

To discuss the first of these subsets first, I suggest it is the case that certain things obviously are or are not reasons for certain dispositional changes.

For example, I suggest that a picture of a logo overlaid over an image of beautiful countryside (to which the organisation signified by the logo has no connection) doesn't represent or present any reason to preferentially deal with the organisation signified by said logo. Nothing in such a picture recommends in any way preferentially dealing with the organisation signified by the logo, even though the picture may cause preferential dealing through some sort of positive association[272]. I suggest, further, that this absence of reasons represented in the picture is obvious. It would be a brave account of the distribution of reasons that tried to deny it in any ordinary circumstances.

Where an influence fails to represent any reasons for change in the dispositions that it influences on such obvious grounds it can be safely concluded that it fails my test of legitimacy, for all respectable accounts of what's a reason for what for whom will concur that the influence fails to represent relevant reasons.

It must be remembered here that whether-or-not something is a reason isn't trivial. As Scanlon said[273] in the right circumstances anything can be a reason for anything, and furthermore anything can be a reason for any given change in dispositions given the right context (as I noted in §3.11). The ubiquitous actuality of reason-hood doesn't follow from the ubiquitous possibility of reason-hood, however. Particulars of a thing and its context determine whether said thing counts as a reason for certain changes in behaviour, similar particulars determine that other things don't count as reasons

[272] D.Langton (2011), *Visual Marketing*, Wiley, New Jersey; pp.23-24
[273] T.M.Scanlon (2014), *Being Realistic about Reasons*, Oxford, Oxford University Press; p.30

in the same way[274]. The taste of ice cream to me – as determined by its chemistry and my physiology – determine whether this taste constitutes a reason for me to establish a disposition to eat more ice cream. Jupiter might be a certain shape, but this shape, it turns out, is unconnected to my ice-cream eating such that it fails to constitute a reason for me to become disposed to eat more ice cream. Being a reason for a change in dispositions is not a trivial property but rather something which depends upon a thing's nature and context.

This non-triviality is important because, at certain extremes, the force of reasons can become vanishingly small. The fact that the Eiffel tower is made of over 7000 tons of iron may give one very little – but some – reason to become disposed to try and eat the building, iron being an important nutrient[275]. Similar 'little reasons' seem to be ubiquitous and can be spotted easily with respect to many things and dispositions. The ubiquity of little reasons, once accepted (one could instead propose there are no such 'little' reasons[276]), might lead one to regard reason-hood as trivial. This could cause problems for the rationalizability requirement, as this requirement uses the effective presenting or representing of reasons to legitimate influences. Agreement about the legitimacy of influences could become meaningless given the ubiquity of little reasons, as reasons would be everywhere and most influences with any representational or presentational content could be legitimated by some little reason somewhere. Agreement about even 'obvious' reasons might become lost in a sea of too easy legitimacy, rendering the rationalizability requirement unworkable. Call this the 'triviality objection'.

It's not obvious to me that little reasons are indeed ubiquitous. Very many things seem to not be even little reasons for given dispositional changes. Even if little reasons are indeed ubiquitous, however, this cannot trivialise reason-hood in such a way as to make the rationalizability requirement unworkable.

Being a reason for something may come in a variety of forms. Some reasons are sufficient; in themselves they warrant belief or – in practical contexts – dispositional change. Other reasons are insufficient; they may become part of sufficient conjunctions of reasons, but on their own they cannot make it the case that an agent ought to make a change in belief or dispositions. What I've called 'little' reasons are systematically of an insufficient sort. Indeed, they're *very* insufficient: they require combination with much additional rational force to alter what ought to be done.

The rationalizability requirement demands that legitimate moral-dispositional influences show reasons for the changes they cause, but not that they necessarily show sufficient reasons for these changes. This allows the triviality objection to threaten the rationalizability requirement, insofar as it allows even little reasons – which seem like they oughtn't make us do anything (at least on their own) – to legitimate influences which work by showing them.

I allow showings of insufficient reasons to legitimate influences because demanding that legitimate influences represent sufficient reasons for the changes in dispositions that they cause would necessarily require demanding that there *be* sufficient reasons for such changes in dispositions. This precludes being legitimately influenced to make a dispositional change one shouldn't, as it would follow from this that one may only be legitimately influenced to make changes that there exists sufficient reason to make (it cannot be the case that there exists sufficient reason for one to behave

[274] *ibid.* p.29
[275] M.Schroeder (2007), *Slaves of the Passions*, Oxford, Oxford University Press; pp.95-96
[276] *ibid.* pp.86-87

in a way in which one shouldn't). This sacrifices the intuition[277] (defended in §3.7) that agents may sometimes be legitimately (in a pro tanto sense) influenced towards doing what they – sui generis – shouldn't. A tyrant murdering my friend, say, seems like a legitimate influence towards rebellion even if, overall, reform would be better than rebellion. At least, said murdering seems like an unproblematic influence on one's rebelliousness in a way that a neurological intervention intended to make one rebellious doesn't. Requiring sufficiency from the presented or represented reasons which legitimate influences would sacrifice such distinctions.

On the other hand, just barely 'being a reason', being a 'little reason', doesn't seem prima facie enough for a thing to be a legitimate influence on behaviour, at least given the triviality objection. *That* the Eiffel Tower is made of iron may count as a reason to eat it, but from this does it follow that a representation of the iron content of the Eiffel Tower is a legitimate influence on dispositions towards eating the building?

Yes.

If someone is moved by a depiction of the iron content of the Eiffel Tower in such a way that he attempts to eat the building, the problem isn't that he isn't moved by a reason but rather that he is moved wildly out of proportion to the strength of said reason. This could either be because the influence is augmented by the addition of irrelevant influential content – something my account objects to[278] – or could be because the character wildly overreacts to the reason he is shown. Where the case is one of pure overreaction (one where the influence only shows a reason that is mis-reacted to) I suggest we should – if appropriate – blame the influence-subject for reacting badly but not call the influence illegitimate or, depending on the circumstances, the influencer unethical.

Suppose the source of the influence that leads someone to attempt to eat the Eiffel Tower is a purely informational leaflet, which states 'the Tower contains over 7000 tons of iron'. It would be bizarre to regard this leaflet as a problematic influence even if it led somebody to attempt to eat the building, and thus blame the leafleteer for influencing behaviour in a bad way. It would make more sense to regard the would-be Tower-eater as somehow defective – like a young child who lacks the capacity to respond properly to reasons (where 'responding properly to reasons' is understood functionally as what one does when one reliably reacts to reasons more-or-less like one should) – and visit upon him any appropriate blame or correction. We may also, of course, blame and correct the leafleteer if he knew the Tower-eater was a defective agent and thus failed in some duty of care he owed the Tower-eater, or else if by his leafletting he secured problematic power over the Tower-eater (I'll discuss such power in Chapter 5).

The account of the ethics of moral-dispositional change I've sketched, dependant as it is on the value of connections between reasons and those they're relevant to, cannot be meaningfully applied to those incapable of responding properly (even in a local or temporary way) to reasons. As Sunstein notes, worries about influencing without rational mediation diminish when we influence the irrational[279]. This is because we owe more-or-less rational beings shown reasons in attempting to influence their behaviour, but we need not owe the same to non- or irrational beings such as the Eiffel Tower eater.

---

[277] *see* R.Brandt (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318; pp.312-313

[278] see §3.9

[279] C.Sunstein (2016), *The Ethics of Influence*, New York, Cambridge University Press; pp.140-141

This point should only be taken so far. Some defectiveness in reason-responding is ubiquitous[280] (though genuine defects in reason-responding should be distinguished from merely apparent defects which may occur whenever subjects are shown insufficiently vivid, or obscure, reasons) but nonetheless I do think we're often obliged to influence moral dispositions only by showing relevant reasons. I suggest that only where this defectiveness is *serious* – as with the child still living amidst the fantasies of youth or the Tower-eater – should we take ourselves to shed the normal duties of influencing. Such serious defectiveness in turn may be defined by a functional disconnect between reasons shown to an agent and their dispositional changes, such that the latter is greatly out of proportion to or wildly divergent from the recommendations of the former.

Where the triviality objection errs is in expecting that an account of the legitimacy of influences prevent little reasons from making any difference to dispositions. In any normal agent such little reasons should be prevented from making any difference to dispositions by their very littleness and consequent inability to move any more-or-less rational agent. Once the weakened triviality objection is stripped of its opposition to little reasons wielding any legitimate influence, however, the form of triviality it cites ceases to be a problem. True (perhaps), little reasons are everywhere. True, my account allows these little reasons to legitimately (in my pro-tanto sense) influence behaviour. This influence, however, must be in proportion to the littleness of the reasons in question. Where it isn't – where agents overreact to such little reasons – the agents in question are either non- or irrational such that they're not governed by the rationalizability requirement, or else are victims of exactly the sort of illegitimate influencing I reject.

We can agree, then, about the legitimacy of influences that obviously show or don't show reasons, safe in the knowledge that our agreement won't be rendered meaningless by any triviality afflicting reason-hood.

What, though, may we say about influences which don't show anything whatsoever to those they influence? Such influences comprise a second class of influences that I think may be ruled pro-tanto illegitimate (and hence such that they must be justified or else resisted) given all views about what's a reason for what for whom.

A physiological intervention, say, (like a drug) intended to effect dispositions can work in multiple ways.

It may, say, alter the senses of the agent such that they're more able to perceive and respond to the reasons they encounter (some theorists describe the action of oxytocin on altruism like this[281]). A person who has lost her sense of smell might perceive no reason to preferentially consume chocolate but find such reason in the food's enjoyable taste once her smell is medically restored. In such a case the smell-restoring treatment seemingly influences behaviour by making a reason apparent to the agent; by enabling the presentation (or representation) of the reason. Provided this reason is genuine (provided chocolate's enjoyable taste counts in favour of preferentially consuming chocolate) per my account the physiological intervention in question is not a pro tanto legitimate influence on the agent's chocolate-eating *in itself*, but rather enables such an influence.

---

[280] V.Milevski (2017), 'Weakness of Will and Motivational Internalism', *Philosophical Psychology*, vol.30, no.1, pp.44-57; p.45

[281] N.Levy, T.Douglas, G.Kahane, S.Terbeck, P.J.Cowen, M.Hewstone and J.Savulescu (2014), 'Are You Morally Modified? The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-171; p.117

'Influence enablers' of this sort may be distinguished from influences proper inasmuch as they don't themselves influence moral dispositions but rather dispositions for responding to moral-dispositional influences, whether these influences work by showing reasons or not. In the chocolate case, the influence-subject will still have to taste chocolate for their dispositions to shift. With oxytocin, if it's an empathy enhancer, the influence-subject will still have to empathise with another's experience for their dispositions to shift. In a case where somebody is rendered highly suggestable by a stroke, the influence-subject will require a suggestion for their dispositions to shift (not all influence enablers of this sort need be 'physiological interventions' – one might suggest certain language skills, say, as members of this class in appropriate circumstances).

Generally, where such enablers occur it should be asked whether the influences they enable are pro-tanto legitimate by my standards; do they work by showing their subjects reasons for the dispositional changes they cause? If such enablers themselves must be evaluated, I suggest they should be evaluated among other things according to how much they facilitate or frustrate the satisfaction of the rationalizability requirement by influences. If a stroke, say, makes one suddenly sensitive to genuine reasons to give to charity, reasons one previously missed (being, say, too jaded to notice them) this gives us a pro-tanto reason to tolerate the stroke. If the same stroke renders one reactive to some non-reason in a way that might disturb one's moral dispositions (makes one more vulnerable to the sort of rhetoric that moved Bernard in AB, say), then this gives us a pro tanto reason not to tolerate the stroke. In general, when we ask whether a particular influence-enabler is good or bad, we should consider (alongside other things; some enablers will be expensive, say) and evaluate the range of facilitations or frustrations of the satisfaction of the rationalizability requirement that the enabler generates[282, 283].

If a physiological intervention is to be an influence in-itself, however, it must function by causing the dispositional changes it causes without first facilitating the showing of anything to the agent.

Many influences on dispositions work this way. For example, I might print a leaflet in a very pretty colour. This colour is simply in-itself beautiful; it doesn't *say* anything (it is not a colour that, say, carries conventional or intrinsic meaning[284]) but is simply such that people find it pleasant to look at. Suppose that much of the influential effectiveness of the leaflet results from its use of this colour. The leaflet carries a message – true – but this message, on its own, would barely motivate. Mostly through the use of the pretty colour in the decoration of its message does the leaflet effect dispositions. In this case the thing that makes the leaflet effective, the pretty colour, neither represents nor presents anything. This thing – this use of colour – is thus, per the rationalizability requirement, an illegitimate influence upon behaviour. It follows from this that the portion of the

[282] It might sometimes simplify thinking to merge what I've called influence enablers into influences by considering them in conjunction with influences. Smell repair is an influence enabler. Tasting chocolate is an influence. Smell repair followed by tasting chocolate is, taken together, both, and hence an influence. The claims of the rationalizability requirement are neutral on such thinking; what matters with respect to the requirement is still only that one's dispositions are changed by relevant reasons. If you can only perceive these relevant reasons because an influence by its action made you able to perceive them this makes no difference here. Similarly, if an influence both makes one vulnerable to the influential effects of some non-reason and shows one said non-reason, this influence no less violates the rationalizability requirement with whatever effects it has on one's dispositions. Generally, there should be no case where discussing things in this conjunctive way (or not) should make any difference in assessments of whether a bad kind of influencing has been deployed.

[283] Within a pejorative language of manipulation, Noggle proposes a similar account of what he calls 'salience nudges'. *See* R.Noggle (2018), 'Manipulation, Salience and Nudges', *Bioethics*, vol.32, no.3, pp.164-170; p.168

leaflet's influence that resulted from the use of the pretty colour (including its use in conjunction with the leaflet's message, distinct from the effects of the message itself) was pro-tanto illegitimate and thus that the overall leaflet, while not necessarily an *entirely* illegitimate influence (it still said something after all, perhaps something relevant), was nonetheless very dodgy. Other considerations aside, we thus ought to oppose the effects of the leaflet on dispositions, by ignoring it, not distributing it, trying to ignore the pretty colour if we do read it, etcetera.

There are complexities to such cases, of course. For example, it may be difficult to distinguish between a case where one is drawn to the leaflet by its pretty colour but only influenced by its representational contents (such that the effect of these contents upon one would have been no different were one's attention drawn to the leaflet in another way) from one where the way one responds to these contents is distorted by the prettiness of the colour. In such a case if the former then the influencing in question may meet the rationalizability requirement; if the latter then this will not be the case.

Furthermore, humans assign meaning liberally enough that proving an influence shows nothing whatsoever will often be difficult. There are few colours as meaningless as the 'pretty colour' suggested above[285].

Absences of effective presentational or representational content in influences are more easily shown in cases where dispositional change is achieved without perceptual mediation, notably through physiological intervention (though some accounts of Christian 'grace' suggest appropriate mechanisms[286]). When Phineas Gage sustained his famous brain injury[287] and was reputedly rendered disinhibited relative to his previous self[288], it was not because the steel rod that shot through his head somehow showed him some fact which challenged his inhibitions. Rather, the injury directly caused the dispositional change by acting on parts of his physiology, without any mediation by his perceptual apparatus. The influence upon Gage, his accident, plausibly didn't change him because it showed him anything, never mind any relevant reasons.

If something bypasses perception when effecting dispositions, it can't be doing so by showing anybody anything (including any reasons) and can't comply with the rationalizability requirement. Hence all such methods of changing dispositions, used to effect change in morals, are ruled illegitimate by my account and must be instrumentally justified or else opposed. This remains true whatever is a reason for what for whom, and whatever our disagreements about reasons.

Furthermore, where we have difficulty saying whether an influence shows anything to its subject, I suggest we gain some reason to regard said influence as illegitimate in this sense. This happens in cases where the representational or presentational contents of influences are *unclear*; where whatever reasons they might show are somehow obscured so that we cannot be very sure whether they're shown at all. Some clarity seems necessary for successful presentation or representation (if it's sufficiently unclear what you're showing, you aren't showing anything), and more than just *attempting* to show some relevant reason is needed to legitimate an influence. It needs to be done,

---

[284] A.K.Fetterman, M.D.Robinson and B.P.Meier (2012), 'Anger as "Seeing Red': Evidence for a Perceptual Association', *Cognition and Emotion*, vol.26, no.8, pp.1445-1458; p.1446

[285] S.Won and S.Westland (2017), 'Colour Meaning and Context', *Colour Research & Application*, vol.42, no.4, pp.450-459; p.451

[286] K.A.Rogers (2004), 'Augustine's Compatibalism', *Religious Studies*, vol.40, no.4, pp.415-435; p.421

[287] M.Macmillan and M.L.Lena (2010), 'Rehabilitating Phineas Gage', *Neuropsychological Rehabilitation*, vol.20, no.5, pp.641-658; p.642

[288] *ibid*. p.642

for this showing must be responsible for the influence's effects and something absent cannot assume such responsibility. Thus, where clarity lacks the chances of legitimate influencing diminish. These diminished chances give us reason to treat influences whose presentational or representational content is unclear as pro-tanto illegitimate.

In summary then, there seem to be two sorts of cases where even those who disagree about what's a reason for what for whom may find agreement about influences' legitimacy. Such opponents may agree with each other about the pro-tanto legitimacy of influences which present or represent only things which are obviously reasons or non-reasons. They may also agree that influences which fail to present or represent anything – or are sufficiently unclear – are pro-tanto illegitimate. I'll show in Chapter 6 how this space for agreement may ground critique of practice.

(4.6) Moral Dilemmas and The Ethics of Moral-dispositional Change

We now come to the second major problem with applying my account: moral complexity. This a problem of balances. How can we weigh the demand of the rationalizability requirement, that we ought to ensure that only reasons influence moral dispositions, against other things we care about? Where we must decide whether to tolerate the influence of something irrelevant, for the sake of realising some other good (say), how should we weigh preventing illegitimate influence against the second good, whatever it is?

We must think in such terms. The account I've offered is a moderate instrumentalism. The reason for this is that the rationalizability requirement is not, I suggest, sufficient to ground an inviolable duty to ensure that influence only proceeds in particular ways. We do – I argue – care about how people's moral dispositions get changed, but there are limits to this caring and features to this caring that have nothing to do with the rationalizability requirement (we might prefer cheap influences, say, as well as rationalizable ones). We can see this when we consider the many cases where illegitimate influencing by the lights of the rationalizability requirement in the service of some greater good seems justified. We might think of a life-ruining addiction cured by a medication[289], say. In such cases we seem willing to tolerate methods of influencing we find problematic for the sake of this-or-that greater good.

If this is no mistake (and I don't think it is), this suggests to us that there are circumstances where the rationalizability may be overridden. In which circumstances, then, should we work to ensure influence proceeds by legitimate means and in which circumstances should we instead maximise or protect other values?

There's no comprehensive answer to this question which can be offered here. As with analogous questions asked of other partial accounts of ethics (consider the banking ethics example from §2.4), answering this question fully would involve adopting a comprehensive ethical theory. Nonetheless, I think there are some helpful general claims about the relations between the rationalizability requirement and other things we value which I can offer.

The first (and least problematic) general claim states that where the choice of method of moral-dispositional change makes no difference to anything else of value one should ensure that only methods which work by showing reasons are employed.

Suppose I find myself struggling with an addiction which inflicts mayhem on myself and those around me. Plausibly the dispositional changes involved in losing such an addiction are important in

[289] X.Koenig and K.Hilber (2015), 'The Anti-Addiction Drug Ibogaine and the Heart: A Delicate Relation', *Molecules*, vol.20, no.2, pp.2208-2228; p.2222

the special way that we flag by calling them ‘moral’. Suppose that, as it happens, I could shed my addiction and achieve this moral-dispositional change either by observing the negative effects of my addiction and being horrified into changing my ways, or by taking a drug which selectively modifies the cells in my brain which maintain my addiction such that I‘m rendered no longer addicted. Suppose that these ways of removing my addiction are in all instrumental respects identical; they’re equally effective, safe, informative (suppose neither method tells you anything you don’t already know), reliable, harrowing, costly and so on, and it’s not the case that their combination would somehow work better.

I suggest that if the rationalizability requirement says anything it says that, despite their instrumental similarities, we shouldn’t be indifferent between these proposed methods. The first method plausibly works by the person undergoing moral-dispositional change perceiving reasons to make the change they make; the second method doesn’t work in such a way. Given such a finely balanced choice, I suggest, we ought to select the first method instead of the second. This point may also be applied interpersonally; intervening in the moral dispositions of another and offered a similarly balanced choice we should select in favour of methods of moral-dispositional change which work by showing reasons. As a conscientious bystander, witnessing the influencing of another in such balanced circumstances, we should object to the use of illegitimate influences and encourage the use of influences which work by the showing of reasons.

We seldom encounter the sort of balanced cases required for this to apply. For one thing, wherever we choose between legitimate and illegitimate methods of influencing moral dispositions we often have the opportunity to combine both methods and create a complicating third. We do however sometimes encounter cases where the balances are unclear, where we do not know the totality of the other values involved or their magnitude. Lacking the capacity to discover, or reliably guess, the weight of these values, an argument may be made for also applying this way of thinking to such cases. If you don’t know and can’t find out which method of influencing moral dispositions will yield the best results, say, or cost the most, and so on, it makes sense to apply the rationalizability requirement to help settle matters.

Such cases are, though, uncontroversial applications of my position. Matters become more complicated when all other things aren’t equal, and the claims distinctive to the ethics of moral-dispositional change must be weighed against other claims on our behaviour. To speak to such (much more common) cases it’s necessary to determine how important the rationalizability requirement is relative to other ethical demands.

The rationalizability requirement is grounded not in a general demand that we be compelled to change moral behaviour only by reasons but rather in the claims these reasons themselves make upon our behaviour. It isn’t the case that dispositions ought to be influenced by reasons and thus we should demand influencing by reasons (though this is a convenient shorthand) so much as it’s the case that specific reasons ought to influence specific dispositions. The rationalizability requirement abducts from the resulting set of individual claims about what ought to shape our dispositions. Behaviour ought to be properly connected to reality, part of this involves having dispositions shaped by the bits of reality that constitute reasons, and what I’ve called the rationalizability requirement captures this.

As the rationalizability requirement is justified not by some general duty to respond to reasons but rather by it being good in many specific circumstances to respond to whatever reasons are relevant, it follows that the strength of the requirement’s demands may vary by circumstance. This is unsurprising. Reasons come in many strengths; it would be strange were the strength of the

demands of the rationalizability requirement unmodified by the strength of those reasons that determine, case-by-case, these demands. Plausibly, the stronger a reason is for changing moral dispositions, the more its role in the changing of moral dispositions deserves the protection of the rationalizability requirement and the requirement's insistence that only reasons should do such changing.

To see how this may help us place the rationalizability in a broader normative context, consider that some reasons seem to influence more important dispositions than others. Some influence us towards or away from thieving. Some influence us towards or away from murdering. This variance in importance is enough, in thinking about what should be done, to make certain reasons outweigh others: reasons not to murder tend to outweigh reasons not to steal whenever the two conflict. We might say, in sui generis terms, that reasons not to murder tend to be stronger than reasons not to steal. It would be odd if this sort of variance in importance didn't make some difference to the demands of the rationalizability requirement. It would be odd if it did not matter more that an agent abide by the rationalizability requirement with respect to reasons concerning more important dispositions which ought to be moved by stronger reasons than less important dispositions which ought to be moved by weaker reasons.

For example, as I noted in §1.1, it seems somehow more tragic for someone to be influenced to rob banks by an irrelevant non-reason than it would be for them to be similarly influenced to buy shoes. Ceteris paribus, it matters more that somebody be moved by shown reasons with respect to his dispositions towards bank-robbing than he be moved by representations of reasons with respect to his dispositions towards shoe-buying. This implies, amongst other things, a certain proportionality between the demands of the rationalizability requirement and the demands of the situation at hand. The more important the dispositions being changed – the stronger the reasons there are for or against such change – the more it matters that change be motivated only by relevant reasons[290].

This axis, I think, determines the general strength of specific demands made by the rationalizability requirement. For each given such demand how much weight one should give the demand (in deciding, say, what to resist the influence of or whether to 'brainwash' by this-or-that method) depends on the importance of behaviour being influenced which is a function of the force of reasons being perceived or ignored. This general strength is what matters, prima facie, in comparing the demands of the ethics of moral-dispositional change to the demands of broader ethics. The more important[291] the dispositions being changed the stronger the relevant demands of the rationalizability requirement.

There are, however, further modifiers on the general strength and character of the demands of the rationalizability requirement.

Numbers count, and the alienation of one's dispositions from relevant reality, like any sort of harm inflicted on only one (particularly, on one's 'reason-responding agenthood'; one's being an agent whose behaviours respond to reasons), can only count for so much. We should plausibly, for

[290] It is also however true that the more important the dispositions being changed the greater the instrumental considerations that must be factored into judgement (its worse for someone to become a murderer than a thief, however it happens). Hence it may be the case that the more important the dispositions the less likely it is for considerations stemming from the rationalizability requirement to outweigh other considerations, *sui generis*. While important dispositions may increase the force of the demands of the rationalizability requirement, it's possible they most often increase the force of other demands *more*.

[291] The limiting cases of unimportance here are entirely non-moral dispositions (see §6.4 for relevant examples) to which I'm not sure the rationalizability requirement applies.

instance, influence tyrants *by whatever means necessary* to prevent them doing evil. Tyrants are few, after all, and this is ethically significant. It makes their fates plausibly less important. Where the proposed subjects of influence are few we're more tolerant of illegitimate influencings, less worried about compliance with the rationalizability requirement. Where the proposed subjects of influence are many, though, this works the other way: illegitimately influencing more people involves harming more people (at least by alienating them from the reasons that should determine their moral dispositions) and this tends to be worse than harming fewer people. Generally, ceteris paribus, better more people be influenced in keeping with the rationalizability requirement than less people. Worse more people be influenced in ways that alienate them from relevant reasons than less people. The strength of the demands of the rationalizability requirement is proportioned to how many peoples' dispositions are being subjected to whatever influencing is at issue.

Furthermore, perhaps justice sometimes demands that certain people be punished[292]. This being the case, you might argue that subjection to illegitimate influence is a fitting punishment for certain individuals; individuals who – somehow – 'have it coming'. Such a 'retributivist' justification of influencing in breach of the rationalizability requirement has been cited (amongst other justifications) in defence of using 'chemical castration' on sex offenders, for example[293]. Such a position may, perhaps, be rendered compatible with my own given a retributivist background theory of justice, though this isn't the place to debate such theories. My account, after all, defines a certain sort of harm in understanding alienation of one's dispositions from relevant reasons as, ceteris paribus, a bad thing to happen to one. Retribution is about inflicting deserved harms and there's no prima facie reason why the harm of having one's dispositions alienated from one's reasons needn't be a deserved response to certain wrongdoings, though this would have to be proved.

In practical contexts, furthermore, there's usually a further modifier on the import that the concerns generated by the ethics of moral-dispositional change should be assigned. Methods of influencing moral dispositions may be more, or less, effective. We may hypothesise methods which are truly effective, and guarantee specific dispositional impacts, but actual methods are hit-or-miss. Some means of influencing work on some people but not others. Some work on some dispositions but not others. Some generate only short-lived, fragile dispositional changes. Some simply don't change dispositions very much.

This variability in the effectiveness of influencings of moral dispositions, I think, implies an additional variability in our evaluations of moral-dispositional influencings. The harm of changing an agent's moral dispositions through an illegitimate means consists not in your mere *trying* to do so but in the actual effects that you have on said dispositions and the agent's consequent diversion from a better sort of agenthood. True, merely *trying* to achieve such things is nasty, in a way, but what matters with respect to whether illegitimate influencing actually happens is what is actually achieved. From this it follows, I think, that more effective means of influencing moral dispositions – those more likely to achieve change – should be taken more seriously. This applies both when the means of changing moral dispositions in question are illegitimate and when they are legitimate. The more compelling an illegitimate influence is the more justification one needs to employ it. The more compelling a legitimate influence is, the more justification one must have to acceptably work to limit its effects upon dispositions.

---

[292] M.N.Berman (2013), 'Rehabilitating Retributivism', *Law and Philosophy*, vol.32, no.1, pp.83-108; p.85

[293] T.Douglas (2014), 'Criminal Rehabilitation through Medical Intervention: Moral Liability and the Right to Bodily Integrity', *The Journal of Ethics*, vol.18, no.2, pp.101-122; p.105

Thus, in evaluating the strength of any given demand of the rationalizability requirement with respect to the ways in which you influence or respond to influence, you should consider three things. First, you should consider how important the dispositions are that are subject to influence. The more important, the stronger the demand. Second, you should consider how many are to be influenced by the influencing(s) the demand pertains to. The more, the stronger the demand. Third, you should consider how effective the means of influencing to be employed as part of said influencing are. The more effective said means, the stronger the demands of the rationalizability requirement.

This rough calculus tells us how we should, generally, evaluate the demands of the rationalizability requirement in considering what we should do. It turns what would otherwise be a flat requirement into a complex normative terrain of weaker and stronger constraints on ethical conduct. It doesn't yet say how these constraints should be weighed against other ethical constraints.

How many perfectly effective illegitimate influencings pertaining to the moral dispositions around killing innocent bystander Greg may be justified in order to, say, save Greg's life? Plausibly, we might think it acceptable to (per the rationalizability requirement) illegitimately influence one person to save Greg. It's bad to manipulate someone's dispositions without showing them reasons, sure, but plausibly less bad than letting Greg be killed (provided allowed harms count for something). We might also be happy to illegitimately influence two people to protect Greg's life, indeed three or four. Eventually, though, I suggest, as the numbers grow higher, we'll grow less sure in our conviction that the illegitimate influencing involved is justified to save Greg. It would seem at least a more troubling dilemma to have to choose between influencing ten million people in unrationalizable ways and allowing Greg to be killed (presuming that the ten million don't somehow 'have it coming'). It would seem wrong to illegitimately influence the entirety of the human species into changing their dispositions pertaining to killing Greg to protect Greg's life. A single human life is very important, of course, but not I think worth harming the entire human species by alienating them from some part of what ought to determine their behaviour. The value of being moved by only relevant reasons is not so vanishingly small as this would require.

This implies that the saving of a Greg's life is worth tolerating something between the illegitimate influencing of nobody and the illegitimate influencing of everybody. It's worth some of what we might call 'brainwashing' (in §6.3 I'll detail how the rationalizability requirement interacts with this concept), but not universal brainwashing, or even perhaps the brainwashing of a great multitude. It would be folly to attempt to quantify how much illegitimate influencing the protection of Greg could justify – for one thing such a number might even depend upon facts about Greg beyond his innocence and bystander-hood (does he go on to save ten lives in a boating accident?). The point here is, though, that there is such a number and that it seems large but finite. This should be enough to show us something about a wide variety of dilemmas we might imagine wherein Greg's life comes under threat and may be saved by illegitimate influencing. Were Greg stalked by a small gang of assailants, for instance, I suggest we'd unproblematically influence them to end their pursuit of Greg with whatever means are at our disposal.

In practice, however, before such considerations come into play it's often the case that the character of the rationalizability requirement itself prevents there being a true moral dilemma between fulfilling this requirement and maximising whatever good may be maximised by influencing.

This may be because the goods competing with the requirement lack moral force.

It may be pragmatically very useful to me if you vote Purple, and one way in which I may ensure that you vote Purple may be by taking measures to reconfigure your moral dispositions such that you vote Purple. Purple candidates might favour subsidising my business, say, and it might be the case that by influencing your moral dispositions in certain ways I can make you vote Purple. To make you vote Purple (and let's say that, here, voting Purple is a moral matter; it's either morally good or morally bad) I could, say, illegitimately influence you by presenting you with compelling non-reasons to change your ways. I could, for instance, work to ensure that the colour purple, which signifies the Purple movement, carries a strong positive association for you and in this way render you positively disposed towards voting Purple[294]. In doing such a thing I may justify myself by pointing to my pragmatic reason; it's of practical importance to me that my company be subsidised, and thus I think that I am justified in seeking otherwise illegitimate influence over your behaviour. I take it that my pragmatic consideration overrides any claims made by the rationalizability requirement. Any justification like this must fail, however, for it must ignore either the moral quality of the rationalizability requirement or the distinctive overriding quality of the moral.

The ethics distinctive to moral-dispositional change I've presented and glossed with the rationalizability requirement is an ethics, not merely a source of advice. As moral reasons the pro-tanto reasons supplied by this ethics of moral-dispositional change have the special overriding character which I argued in Chapter 2 separates the moral from the non-moral. Specifically, as moral reasons these reasons are sufficient to override conflicting non-moral reasons, no matter how strong they are. Thus, in any cases where the pro-tanto reasons supplied by the ethics of moral-dispositional change stand in conflict with only non-moral reasons these pro-tanto reasons ought to override any other reasons and determine what should be done.

In the example of encouraging voting Purple to secure subsidies for one's business, the influencer thus cannot justify the use of influences declared illegitimate by the rationalizability requirement. If your reasons for influencing are non-moral, and the reasons against using the means you have selected to do said influencing are moral, you do wrong in attempting to influence in the way you have selected.

More vexing dilemmas may be had where there is a moral reason to employ illegitimate influence. Suppose, say, that I wanted you to vote Purple to prevent hardship caused by environmental damage, damage more likely to be prevented were you to vote Purple. Preventing such hardship seems to be a moral matter in a way that being subsidised is not (let's say – though of course it might be, if for instance livelihoods are at stake). Where the choice is between preventing such hardship using illegitimate influence and failing to prevent said hardship in an effort to avoid influencing illegitimately we seem to have a true moral dilemma.

Such moral dilemmas depend on the existence of a dichotomy which only occurs whenever specific conditions obtain. You can only have such a dichotomy between the demands of a moral reason and the demands of the rationalizability requirement whenever you're somehow unable (say due to possibility constraints) to secure the moral-dispositional change you seek by showing the moral reason itself.

Where the rationalizability requirement demands that you show somebody S a reason to change their moral dispositions, and broader ethics demands *that* you change S's moral dispositions, prima

[294] I.Kareklas, F.F.Brunel and R.A.Coulter (2014), 'Judgement is not Colour-Blind: the Impact of Automatic Colour Preference on Product and Advertising Preferences, *Journal of Consumer Psychology*, vol.24, no.1, pp.87-95; pp.91, 93

facie you ought to have some chance at satisfying both requirements by showing S the moral reason for changing their dispositions.

For example, it might be morally important that you give to charity. If it's morally important that you give to charity, it will be so on account of some moral reason. Perhaps there's some suffering out there that you ought to alleviate, and which could be alleviated by your charity. In such a case to secure a change in your moral dispositions such that you give to charity, I (or any other agent interested in changing you, perhaps including yourself) ought to be able to show you the reason that makes it the case that you ought to give to charity. I could show you the suffering you ought to alleviate, for instance. In doing so I could both satisfy the demand of broader ethics, that you be influenced to have certain moral dispositions, and the demands of the rationalizability requirement, that you be influenced only by presented or represented reasons.

In practice, however, one cannot always satisfy the demands of broader ethics and the demands of the rationalizability requirement simultaneously in this kind of way. This is the case even though it is the case that wherever these demands come into conflict the demands of broader ethics themselves furnish the influencer with the kind of reasons needed to satisfy the demands of the rationalizability requirement.

In practice there are many defects of communication and agents that can make it difficult, costly or even impossible to influence by showing reasons, even where reasons are available. The influencer and the subject of influence may, say, lack a common language and it may thus be difficult for the influencer to verbally show the subject anything. The subject of influence may hate the influencer so much that he will resist any influence the influencer deploys as a matter of course. The influencer may have very little time to influence the subject of influence, not enough to vividly portray whatever reason the subject has for changing moral dispositions.

In such cases, however, I suggest that apparent dilemmas between the demands of the rationalizability requirement and the demands of broader ethics are very often rather trilemmas between the demands of the rationalizability requirement, the demands of broader ethics and the costs involved in satisfying both.

Where a common language is lacking, and reasons cannot be represented, a common language may be found or created or else reasons presented in other ways. Where hatred blinkers perceptions, an alternate vector of influence can be found. Where time is short, time can be made by the clearing of other business. In practical circumstances, the question is only rarely 'should I breach the rationalizability requirement or sacrifice broader ethical concerns?' It is more often 'should I breach the rationalizability requirement, sacrifice broader ethical concerns, or pay the price necessary to do neither?'

A true dilemma between the rationalizability requirement and broader ethics thus only occurs when it's somehow impossible to change the situation such that influence may feasibly be conducted without sacrificing anything morally important.

Such a dilemma may occur if the proposed subject of influence themselves is somehow defective in their ability to respond to reasons, like the putative Eiffel Tower eater I mentioned earlier or a young child. I've suggested that the rationalizability requirement cannot be applied to such radically non-rational beings. Influencing these beings involves other standards. Any agent who isn't radically defective in their ability to respond to reasons will be such that they may be influenced by whatever reason grounds the broader ethical demand that they have their moral dispositions changed (at least

to some extent). Such a responsive agent may nonetheless be costly to influence by such reason-showing, however, and cheaper to influence using illegitimate methods.

Where the costs involved in influencing someone legitimately lack moral character (where they involve, say, only inconvenience), I suggest, they ought to be paid in keeping with the overriding nature of moral considerations (see §2.2). If it is morally important that someone be influenced, and morally important that they be influenced by a reason, and not morally important that it is – say – time consuming to influence using said reason, the right thing to do is influence using said reason. Where the costs involved in influencing someone legitimately have moral character, on the other hand[295], we are faced with a moral trilemma. We must either sacrifice the values captured by the rationalizability requirement, the values captured by other ethics, or whatever values must be sacrificed as part of the cost of legitimately influencing. These lattermost costs need not be proportional to the other sets of costs and thus nothing in general may be said about them. If they're great enough then sacrificing either the values captured by the rationalizability requirement or the values captured by other ethics in play will be warranted, otherwise not. It is in such cases that we can face what seem to be dilemmas between the demands of the rationalizability requirement and other demands.

Consider Jill, who ought to stop being racist, and ought only to be moved such that she stops being racist through exposure to some reason implying that she should stop being racist. Suppose that, for some reason, satisfying both of these 'oughts' by influencing Jill using the reasons that make it the case that she ought not be a racist would involve some morally onerous sacrifice. Perhaps the reasons in question could only be communicated to Jill if she is taught a language in which they are expressed, and this would involve a large bill which would have to be paid at the expense of completing some morally valuable project. Nonetheless Jill is, let's say, a proper responder to reasons; she isn't defective like the young child or Eiffel Tower eater and thus beyond the aegis of the rationalizability requirement. In such a case, if Jill ought to not be a racist and ought only to be moved against racism by a relevant reason, these demands cannot be simultaneously satisfied. We might be able to, say, satisfy the first demand by applying some illegitimate method of moral-dispositional change, but in doing so we would have to ignore the demand of the second claim on what should happen to Jill. On the other hand we could satisfy this second demand at the expense of the first, forgoing the illegitimate influencing of Jill in the name of not influencing her in a wrong sort of way at the expense of leaving her racism intact.

What we ought to do, given these choices, will depend on the particulars of Jill's case and their importance. For example, what are the consequences of Jill's racism? If they're relatively minor, consisting only of (say) occasional cruel words, the reasons these consequences generate need not be strong enough to override the requirement that Jill only be influenced by relevant reasons. If they are relatively more important though, consisting of (say) participation in racist violence, then these consequences are plausibly more likely to generate sufficient reasons to overrule any claims of the rationalizability requirement.

What is important here is that such overruling can only occur in certain circumstances. It can only occur whenever the moral cost of satisfying both the rationalizability requirement and broader ethics is high and the relevant broader ethics can overrule the claims of the requirement. In other contexts where legitimate means are ethically cheap or the claims competing with the

[295] As one might suppose they always do, if – say – the resources spent on legitimately influencing might always be alternately committed to some morally valuable project. This might but need not be the case; hence I allow for the possibility of purely non-moral costs here.

rationalizability requirement's claims are relatively minor the requirement can at least partly determine right action.

(4.7) Conclusions

I have attempted to show how my account may still yield insight into the rights and wrongs of influencing despite competing ethical claims and disagreement (perhaps uncertainty) about the distribution of reasons. However reasons are distributed we can agree that some things are or are not reasons for some agents and evaluate influences accordingly. This agreement is rendered meaningful by the non-triviality of being a reason. Things that present or represent nothing (or present or represent in sufficiently vague ways), furthermore, cannot meet the demands of the rationalizability requirement, whatever is a reason for what for whom. I suggest that dilemmas between the rationalizability requirement and other ethics are possible. Sometimes, however, they will turn out to involve agents to whom the rationalizability requirement cannot be applied. Often, also, such dilemmas may be dissolved by employing representations of the reasons grounding claims competing with the claims of the rationalizability requirement as influences. Some such apparent dilemmas will also lack moral depth. Where insoluble moral dilemmas between the rationalizability requirement and broader ethics occur, these must be judged case-by-case. In such judgement the strength of the demands of the rationalizability requirement will be proportional to the importance of dispositions undergoing change, the number whose treatment these demands pertain to and the effectiveness of moral-dispositional influences involved.

Chapter 5: Freedoms and Moral-dispositional Influencing

(5.1) Abstract

In this chapter I discuss freedom. I try to show that if one values freedom – whether one understands it as autonomy or as non-domination – one ought to accept the rationalizability requirement. If one values freedom as autonomy then one must commit to one of two ideals concerning agency, one of which supports the rationalizability requirement while the other doesn't. The ideal which doesn't support the rationalizability requirement, however, is vulnerable to charges of 'moral fetishism[296]'. This being the case, someone who believes autonomy considerations should determine our evaluations of influences has reason to embrace the rationalizability requirement. If one values freedom interpersonally as non-domination, furthermore, one has reason to embrace the rationalizability requirement to ensure moral dispositions are influenced by non-person reasons rather than potentially domineering other people. This is the case even though such others can curate the representations of reasons that influence us. Given all this at least two major kinds of proponents of freedom have reason to embrace the rationalizability requirement. This requirement thus has some capacity to unify apparently competing perspectives in extant debate about the legitimacy of influences.

(5.2) Freedom and the Evaluation of Moral-dispositional Influencing

I've argued that we may evaluate influences on moral dispositions by the 'relevance' relations that may or may not obtain between influences on moral dispositions and the moral-dispositional changes they cause. Where an influence is relevant to a change, inasmuch as it shows a reason for it and doesn't owe its effective power to anything but this showing, said influence is of an ethically better sort. Where an influence fails this relevance test something goes wrong and this influence must be justified as a lesser evil.

This thesis and the argument behind it plausibly only tell us part of what we need to know to fully evaluate influences. For one thing, I have arrived at this analysis through examination of cases which are idealised in certain rarefied ways. The case of Annie and Bernard, for example, removed all interpersonal elements prima facie important to our decision making about the acceptability of influences on morals. Plausibly, something we care about in forming our moral dispositions (at least as adults) is that others exert a *limited* influence over this formative process. We might care about this because we believe that moral dispositions that are formed autonomously are somehow better than those whose origins are dependant, or else because we wish to evade others' domination of our behaviour.

We invoke such concerns in asking for 'freedom' to endure through processes of influencing. Conceptions of freedom impinge upon our thinking about the ethics of influencing, including the ethics of influencing morals, routinely. Indeed, they perhaps present the most common extant frameworks in which such thinking takes place[297]. In this chapter I'll investigate how these frameworks relate to my account. I'll show that, whether you think of freedom as autonomy, or as a thing to be secured by escaping the domination of other persons, you should have reason to support my perspective (in the latter case in practice and in the former case at least in practice, arguably in

[296] M.Smith (1994), *The Moral Problem*, Malden, Blackwell; pp.71-76

[297] *see for detail* T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; pp.690-693

principle). Given my objective is to prepare a unifying understanding of the ethics distinctively of influencing moral dispositions, I take this to be a merit of the view I propose.

(5.3) Autonomy and Moral-dispositional Influencing

I'll begin this treatment by addressing arguably the most vocal tradition opposing practices that manipulate moral dispositions in extant literature: the autonomist tradition.

The *Autonomist* about the ethics of influencing morals may hold one of several views. The classic view, popularised by Kant, holds that for actions to have moral value they must proceed from a specific sort of psychology. Specifically, they must result from the outworking of a 'good will', rather than the mere spurring of some sort of incentive such as the need to fit in socially, say, or the desire for happiness[298]. In Kantian jargon, one must choose to adopt the good as a sufficient grounding for one's maxims (those bits of one's psychology that determine one's actions)[299]. The content of the good is dictated by reasons (about which a separate origin story may be told), but what is fundamentally important is that the agent has a 'good will'; that they do their duty because it's their duty. Where behaviour is motivated by something other than the response of the will to the recognition of moral duty, actions are – on this view – morally worthless. If one is only honest in one's shopkeeping because of the need to avoid the ire of one's customers, say, on this view, one's honest shopkeeping lacks moral worth[300].

This way of thinking has a long history. It was popularised, perhaps, by Christ's insistence that to do moral actions for the sake of worldly incentives (such as the approval of one's peers) is to 'have one's reward' in a way that deprives said actions of worth before God[301]. Kant secularises this thought by supplanting the ideal of 'worth before God' with the ideal of the *good will* as 'highest good and the condition of every other[302]'. Kant's intellectual fulcrum in making this move is an ideal of imputability; the idea that only through a distinctive sort of incentive-free willing can the will show its true quality. You can only know a will is truly good, according to Kant, when that will inspires actions which run against non-moral incentives – that is, for Kant, any incentive that isn't supplied by a properly moral motivation[303].

This Kantian 'autonomism' plausibly places limits on the kinds of influences that ought to be allowed to impact people. If what really matters is the good will and this retreats as certain motivations – inclinations – advance[304] then it follows that care ought to be taken to limit the effect of such motivations on moral action. This care seems to be warranted at least in the case of isolated moral actions and plausibly with respect to the tendencies regarding moral action I call moral dispositions. Applying this thinking to dispositions (as one must if one is to offer any autonomist evaluation of

---

[298] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; pp.9-10

[299] I.Kant (1794[1998]) trans. A.Wood and G. di Giovanni (1998), 'Religion Within the Boundaries of Mere Reason' in A.Wood and G.di Giovanni (1998) [eds.], *Religion Within the Boundaries of Mere Reason and Other Writings*, Cambridge, Cambridge University Press; p.49

[300] I.Kant (1785[1998]) trans. M.Gregor (1998), *Groundwork of the Metaphysics of Morals*, Cambridge, Cambridge University Press; pp.10-11

[301] Matthew 6:1-6, Holy Bible: King James Version

[302] I.Kant (1785[1998]) trans. M.Gregor (1998), *Groundwork of the Metaphysics of Morals*, Cambridge, Cambridge University Press; p.10

[303] *ibid.* pp.10-11

[304] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; p.108

methods of moral-dispositional change, per extant literature[305]) yields a challenge to many possible methods of moral-dispositional change, which effect dispositions by creating psychological phenomena which arguably constitute inclinations or 'non-moral incentives'. These phenomena may include things such as sympathies, associations or fears. Such phenomena don't seem to meet the Kantian standards for moral motivations; they're understood instead as non-moral incentives ('inclinations')[306]. These and similar phenomena are, however, important to many methods of changing moral dispositions. Annie from Chapter 3, say, seemingly applied a method which leveraged her sympathy, while the propagandist and advertiser may employ both associations and fears in shaping the dispositions of their audiences. There's a venerable[307] literature drawing upon Kant and Kantians dedicated to opposing all such methods from an autonomist perspective[308].

This being the case it seems that the Kantian autonomist offers a distinct and challenging view of the ethics of influencing moral dispositions, both for my own position articulated in the previous chapter and for alternative positions. I, after all, contend that what Annie did in seeking inspiration in news reports was *okay*, while others (though not I) contend that moral-dispositional changes inspired by associations or medical interventions are *okay*. The Kantian autonomist seems to have grounds to object to all these methods, though. I must thus supply some defence against the Kantian autonomist, at least if I am not simply to dismiss autonomy as an orthogonal value (which I do not believe I should, for as I will argue it seems to be something protected by the application of the rationalizability requirement). I must show that they can and preferably should accept something like my position, at least in practice.

My defensive strategy will involve two moves. First, following existing literature, I'll argue that the quickest form of the Kantian autonomist objection to diverse methods of moral-dispositional change invokes an implausible view of ideal agency[309]. Once this error is corrected the Kantian position gains plausibility but its objections to a wide range of methods of moral-dispositional change must gain nuance. My second defensive move will work by showing that, insofar as the corrected Kantian view is plausible, it is plausible on account of what it shares with my own view offered in Chapter 3. This being the case the Kantian has reason to agree at least with the practical demands of the rationalizability requirement. I'll attempt to show that there is this reason for agreement by highlighting an ambiguity between de re and de dicto interpretations of the form of moral motivation Kant valorised. On the former interpretation, I'll argue, the rationalizability requirement should be endorsed while on the latter it needn't be but this interpretation is itself unworkable.

In offering this defence I'm aware not every autonomist is a Kantian. Indeed, the whole tradition is far too large to treat adequately here[310]. Kant's work, however, supplies the tradition with a *locus classicus*, and most autonomist positions may be helpfully defined in relation to Kantian doctrine. My strategy will therefore be to outline my response to the Kantian position as a helpful proxy for the whole, while addressing some permutations of the ideas at issue and subsequent literature

---

[305] *see for examples* T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; pp.690-693

[306] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; pp.108-109

[307] *see* H.C.Brown (1929), 'Advertising and Propaganda: A Study in the Ethics of Social Control', *International Journal of Ethics*, vol.40, no.1, pp.39-55; pp.44-48

[308] A.Villarán (2017), 'Irrational Advertising and Moral Autonomy', *Journal of Business Ethics*, vol.144, no.3, p.479-490; pp.487-488

[309] *see* A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; pp.30-34

[310] T.M.Scanlon (2011), 'Why not Base Free Speech on Autonomy or Democracy?', *Virginia Law Review*, vol.97, no.3, pp.541-548; pp.546-548

along the way. In this way I mean to speak to as much of the autonomist tradition as possible while at least helpfully locating my claims relative to parts of the tradition I don't address.

(5.4) The Rationalizability Requirement and Kantian Moral Motivation

Kantian autonomists, evaluating influences on moral dispositions, come in two varieties. Let's call them quick and slow. The 'quick' Kantian autonomist is quick because they can quickly object to many methods of changing moral dispositions. They can make these quick objections because they believe that the formation of human moral dispositions is a sort of battleground upon which non-moral incentives, 'inclinations', do battle with the will. This will may be good or bad[311] but either way this will, on the quick Kantian autonomist view, inherently conflicts with inclinations. Importantly this conflict is one of impermeable battle lines. One cannot be motivated to form a given moral disposition by both the content of one's will and one's non-moral incentives; it must be one or the other[312].

This feature of the quick Kantian autonomist's position enables them to quickly reject many methods of changing moral dispositions. They need only ask themselves whether the method of moral-dispositional change at issue employs inspiration by non-moral incentives, that is, anything beyond regarding the adoption of the given moral disposition as a moral duty (which, for the Kantian, is enough to supply a distinctive 'moral' incentive). If so, then the method of moral-dispositional change at issue is problematic insofar as it limits the play of the will and thus precludes action from moral duty.

In trying to commit Kant to this 'quick' position one may cite certain patterns in his writings, notably his repeated use of actions undertaken in the absence of or against non-moral incentives to exemplify what he regarded as properly moral and valuable in human behaviour[313].

The problem with this 'quick' position is that it seems to place bizarre limits on what can count as a moral action. Schiller famously highlighted this by jokingly noting how much of a pity it is that he enjoys his friends' company, since this enjoyment renders his friendship morally worthless[314].

Other thinkers have noted the problem here many times: surely it's possible for the will and one's non-moral incentives to both inspire the same action, and for that action to retain a moral quality? The current received view is that such a thing is not only possible, but a possibility that Kant endorsed. On this view the tendency of Kant's example-set is best explained by his focus on isolating moral from other incentives to make their nature clear[315]. According to this received view, what really matters for Kant is that the moral incentive is itself sufficient to establish a moral disposition, irrespective of whatever other incentives are involved[316]. This, then, is the 'slow' Kantian autonomist position, one which holds that one can't jump to conclusions about the acceptability of influences

---

[311] I.Kant (1794[1998]) trans. A.Wood and G. di Giovanni (1998), 'Religion Within the Boundaries of Mere Reason' in A.Wood and G.di Giovanni (1998) [eds.], *Religion Within the Boundaries of Mere Reason and Other Writings*, Cambridge, Cambridge University Press; pp.55-57

[312] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; p.30

[313] *see for example* I.Kant (1788[1997]) trans. M.Gregor (1997), *Critique of Practical Reason*, Cambridge, Cambridge University Press; pp.72-73, H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; pp.110

[314] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; p.110

[315] *ibid.* pp.110-111

[316] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; p.107, C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; p.58

based upon a preliminary analysis of the psychological levers they pull. You also need, sometimes, counterfactual knowledge about the levers in question.

I'm not sure whether Kant himself was a quick or slow autonomist. I'll merely note that I find the 'slow' reading more compelling, not least because Kant sometimes encouraged the cultivation of certain non-moral incentives (such as sympathy for the poor, by paying them attention[317]) without fretting about the moral qualities of resulting actions. What's more important is that the slow autonomist position is more plausible than the quick one. It seems strange that the over-determination of a dispositional change by a non-moral incentive, such as a desire, should render amoral a disposition that was adopted out of respect for what is right. At least, it seems strange if you are a Kantian and you value adopting dispositions out of respect for what is right as the outworking of a good will. In assessing the potential autonomist challenge to both my position and certain more permissive positions I will thus limit my comments to slow autonomism.

Even without availing of quick autonomism, the Kantian can find fault with methods for changing moral dispositions.

For the Kantian there can be only one proper method for changing moral dispositions: responding to the motivational content of the moral law[318] – what I'll call (proper) Kantian moral motivation. The reason for this is that for the Kantian only moral behaviour resulting from respect or disrespect for the moral law, hence the quality of one's will (preferably good), may be properly imputed to the agent[319]. Anything else is an accident of circumstance and may result in changed dispositions but not dispositions with any worth and hence nothing of value. Roger, say, (from §3.9) who chances upon a homeless person suffering in the snow, may gain new moral dispositions but these dispositions will simply be worthless. They'll be worthless because they aren't motivated by respect for what is right but by the happenstance of Roger's sympathies (they can thus never be more than 'simulacrums' of dispositions of moral worth, as Korsgaard puts it[320]). Dispositions generated in such a way cannot be products of one's will and thus cannot be compatible with one's freedom in an important autonomist sense[321].

This species of objection is available in responding to many of the cases I've discussed, including some which seem to meet the rationalizability requirement. The behaviour of Annie, say, from Chapter 3 – which I accept – is plausibly opposed on this understanding insofar as it betrays her as unable to adopt the right moral dispositions out of pure respect for what's right (she needs to watch news reports to find inspiration to follow her ideals). This being the case prima facie the Kantian autonomists' understanding seems to depart from the rationalizability requirement. It seemingly warns us that that the sort of rationalizable influences I've been valorising can fall short in an important way. They can fail to involve the sort of motivation that proper moral-dispositional influences ought to have. This being the case there seems to be space for such an autonomist to fail to endorse the rationalizability requirement.

---

[317] I.Kant (1797[1996]) trans. M.Gregor (1996), *The Metaphysics of Morals*, Cambridge, Cambridge University Press; p.205

[318] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; pp.111-112

[319] This isn't to say that actions from inclination can't be imputed to agents, but that when such action is imputed to agents it lacks moral worth as it lacks origins properly related to the quality of one's will; *see ibid.* pp.111-112

[320] C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; pp.56-57

[321] I.Kant, M.Gregor and J.Timmerman (2011), *Immanuel Kant: A Groundwork of the Metaphysics of Morals: A German-English Addition*, New York, Cambridge University Press; pp.81, 83, 85, 87

I believe that this Kantian challenge ultimately fails in such a way as to imply that they should rather endorse the rationalizability requirement. I'll attempt to justify this claim by an interrogation of the substance of the moral motivation the Kantian valorises.

There are two ways in which we can interpret Kantian moral motivation. If it is understood as a motivation to do what's right, then this motivation can be understood in de dicto or de re terms. It may involve either a motivation to do what's right, whatever it may be, or a motivation to do those things which are actually right[322]. I contend that it is only when the Kantian moral incentive is interpreted in de dicto terms that it is capable of grounding an objection to my views. I contend further that to interpret Kantian moral motivation in de dicto terms is to deprive it of too much meaning.

I'll begin fleshing out this argument by first addressing the imputability concerns that ground Kant's view of proper moral motivation.

To recap my view: I think that the better methods of moral-dispositional change are those methods which effect moral dispositions only and entirely because of the effects of their showing 'relevant' bits of reality to the disposition-changing agent. I further define the 'relevant' bits of reality as those bits of reality that constitute reasons for dispositional changes in question. To determine whether a given means of moral-dispositional change ought to be accepted[323] one must thus determine whether the means in question works because of the reasons it shows or because of some other factor. If the latter, there is reason to reject the means suggested, not employ it, and strive to resist and encourage the resistance of its influence. There's thus a 'rationalizability requirement' on methods of moral-dispositional change.

According to this view, moral-dispositional changes must ultimately be grounded in the effects of specific parts of reality on an agent or else be found wanting. These parts of reality are in turn specified by reason; they are precisely those parts of reality that count as reasons for the agent to make the dispositional change involved, were the agent to reason about the matter (which they may or may not do). For the Kantian, similarly, the moral law is grounded in the demands of 'practical reason', a species of reason that demands specific actions be taken[324].

Owing to this common grounding it's possible to offer an interpretation of the Kantian doctrine which supports the rationalizability requirement. The Kantian doctrine demands that actions be motivated by the 'moral motivation' or 'moral incentive[325]'. If this 'moral incentive' refers to the normative force of the reasons which ground the moral law, then this 'moral incentive' refers to precisely the force of the reasons which the rationalizability requirement demands inspire moral-dispositional change. As such the Kantian demand that actions result from the incentive supplied by the rightness of the actions turns out to mirror the rationalizability requirement's demand that moral-dispositional changes result from perceiving bits of reality that are reasons for said changes. The Kantian doctrine thus seems compatible with – indeed implies – the rationalizability requirement, at least when applied to the sorts of cases that the rationalizability requirement applies to and the motivation of dispositional changes in addition to isolated actions.

---

[322] M.Smith (1994), *The Moral Problem*, Malden, Blackwell; p.74

[323] at least in certain pro tanto terms distinctive to evaluating moral-dispositional influences

[324] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; p.233

[325] *ibid.* p.122

I commit to this understanding as a response to the Kantian challenge, but before I offer it more carefully in §5.5 I'll briefly address a few initial worries one might have about how I've been articulating the Kantian position.

For one thing, one might worry that my description of the Kantian view of legitimate moral-dispositional change is oversimplified.

Kant divided the will into legislative ('*wille*') and executive ('*willkür*') components[326]. It could be argued that, on a Kantian view, legitimate influencing is not grounded in only practical reason but in the combination of practical reason with the practical reasoning of the *wille*. What's important isn't just responding to the right reasons but also processing these reasons through well-functioning reasoning processes. Roger Crisp, for instance, opposes some advertising methods along autonomist lines on the basis that they subvert reasoning processes[327]. Harris offers a comparable argument in opposing biomedical moral enhancement, arguing that such enhancement done in certain ways destroys the possibility for practical reasoning about what's right[328].

If legislative (*wille*) reasoning as well as responding to reasons is indeed needed for autonomy then any potential Kantian support for the rationalizability requirement is lost. The rationalizability requirement, after all, demands only that dispositional changes be *rationalizable*, not that they be *rational*. It demands that such changes be motivated by reasons but specifies nothing about how these reasons must be processed, *reasoned* about, before they legitimately effect behaviour. If the Kantian view is understood as requiring not just that we be motivated by reasons but that these reasons be processed in some specific way (such as integrated into reflective equilibrium judgements, say) then it isn't clear how this view could ever imply the rationalizability requirement[329]. The requirement would – on such a view – seem to systematically ignore the value of influence-subjects reasoning properly (however this may be accomplished) in response to influences, in separating good influences from bad ones.

The problem with this worry is that it depends upon the muddling of irrelevant – plausibly epistemic – concerns into metaphysics. To support this worry one must hold that what it's right to do is necessarily determined by the conjunction of a set of reasons and the action of some internal reasoning (for only these things together can constitute a legitimate influence on what one does). This seems to me to be confused. True, we often discover what's right by the use of reasoning processes (such as reflective equilibrium judgement), but what's right isn't determined by these processes. What's right is determined – if it's determined by anything – only by facts that constitute reasons, not procedures for discovering them. By being moved by these facts we can be moved by reasons irrespective of what reasoning we do or do not undergo internally.

An opponent could respond here by suggesting that they are not discussing rightness in a general sense, but only in the specific sense in which something may be both right and imputable to an agent. Thus they may argue that motivation by reasons-plus-reasoning is necessary for a change of dispositions to be legitimate *for an agent*, given that the application of reasoning is a pre-requisite

---

[326] *ibid.* pp.129-130

[327] R.Crisp (1987), 'Persuasive Advertising, Autonomy and the Creation of Desire', *Journal of Business Ethics*, vol.6, no.5, pp.413-418; pp.413-414

[328] J.Harris (2016), *How to be Good: The Possibility of Moral Enhancement*, Oxford. Oxford University Press; pp.80-82

[329] save as an insufficient condition on legitimate influencing

for imputability and absent such imputability changes in moral dispositions lose value[330]. Such a response is problematic insofar as by declaring reasoning necessary for imputability it implies that impulsive actions lacking reasoning cannot be imputed to the agent who intuitively did them. On this understanding, acting 'without thinking', say, must be not acting at all, which is a strange position to advocate.

The opponent might also complain that the role of the legislative processing of reasons, reasoning, done by the *wille*, is to decide whether reasons are or are not reasons. Inasmuch as this is the case in demanding we be motivated by reasons-plus-reasoning what my opponent demands is that we be motivated by reasons which we believe are reasons. This is, in effect, a valorisation of de dicto Kantian moral motivation of a sort I'll address in the following section.

Kant's account of idealised agency might also seem to involve at least one further precondition in the form of a spontaneous choice[331] to respond to the motivational power of the moral law[332]. Such spontaneous choices, if they may indeed be made, should be understood as governing our responses to influences (for better or worse) and not the acceptability of influences themselves in any way relevant to the rationalizability requirement. The reason for this is that this spontaneous choice is best understood as constraining what the agent counts as a reason – which is not to be confused with what they *should* count as a reason – and hence what motivates them[333]. The rationalizability requirement, by contrast, demands that we be influenced by those things that we should count as reasons (our 'justifying' reasons[334]), and our spontaneous choices don't in any way determine such things. You can't spontaneously pick what things are or are not reasons for you to X, only whether you respond to them as if they are or are not reasons for you to X.

Given this, per the rationalizability requirement, it's possible for a legitimate influence to retain legitimacy even when stripped of effectiveness by the spontaneous choice of its subject to discount the force of the reasons it shows (presuming such things can happen). This legitimacy would be no greater were this spontaneous choice to go another way and this effectiveness conserved. I might say, given the rationalizability requirement, that stripping a legitimate influence of the motivational power it has in virtue of the reasons it shows by making the spontaneous choice to discount this motivational power (if one can do such a thing) is a pro-tanto bad thing to do, but a Kantian might be expected to agree with me about this[335].

I'll hence limit my comments to what I understand to be the traditional autonomist ideal, conceived of as a form of idealised agency and expressed by Kant. This notion of idealised agency attempts to establish a basis upon which right and imputable actions (explicitly including the adoption of new moral dispositions[336]) can be committed. It does so by singling out actions which are motivated by a

---

[330] I.Kant (1794[1998]) trans. A.Wood and G. di Giovanni (1998), 'Religion Within the Boundaries of Mere Reason' in A.Wood and G.di Giovanni (1998) [eds.], *Religion Within the Boundaries of Mere Reason and Other Writings*, Cambridge, Cambridge University Press; p.46

[331] I.Kant (1788[1997]) trans. M.Gregor (1997), *Critique of Practical Reason*, Cambridge, Cambridge University Press; pp.79-82, M.Sgarbi (2012), *Kant on Spontaneity*, London, Bloomsbury; pp.9-12

[332] H.Allison(1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; pp.129-130

[333] C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; pp.164-166

[334] J.J.Tiley (2004), 'Justifying Reasons, Motivating Reasons and Agent Relativism in Ethics', *Philosophical Studies*, vol.118, no.3, pp.373-399; p.376

[335] C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; pp.166-167

[336] I.Kant (1794[1998]) trans. A.Wood and G. di Giovanni (1998), 'Religion Within the Boundaries of Mere Reason' in A.Wood and G.di Giovanni (1998) [eds.], *Religion Within the Boundaries of Mere Reason and Other Writings*, Cambridge, Cambridge University Press. pp.84-88

special 'moral motivation' and thus escape from the happenstance of action-from-inclination into the realm of truly autonomous action. It is in comparing this motivation with the kind of motivation valorised by the rationalizability requirement that the most direct autonomist challenge to the rationalizability requirement may be explored.

(5.5) Interpreting the Moral Motivation

This direct challenge works by interpreting the good Kantian agent in terms of an ideal motivational structure. The objector stresses the Kantian idea that the moral quality of dispositions depends upon their origins in the proper moral motivations. In order for a change in moral dispositions to be of the good sort, it must be one motivated by the incentive to do what is right, what one has a duty to do[337]. What matters in ethics for the Kantian is why you do what you do, and there can be no better incentive for a change in behaviour than a recognition of the rightness of this change. Indeed, there can be no good incentive for change in behaviour beyond this 'moral incentive[338]'.

This interpretation of the Kantian doctrine produces an objection to my view by then suggesting that the motivation that is supplied by the 'representations and presentations' of reasons demanded by the rationalizability requirement may deviate from the motivation to do what is right. You may, say, be motivated to help someone in need not by the recognition of the rightness of doing so but by mere sympathy. A representation or presentation of the fact of need could certainly do this, if able to evoke sympathy. You may, say, be warded off larceny by recognising someone else's ownership of something, and being moved by their attachment to it, rather than by recognising the wrongness of stealing the thing.

One might understand this objection as accusing my view of an objectionable sentimentalism. Kant was famously suspicious of sentiments, arguing in an exchange with Schiller that they ought properly to be subordinated to the pure practical reason that alone can supply moral motives[339]. The objector may invoke this thread in Kantian thought, arguing that the rationalizability requirement leaves open the problematic possibility of agents being moved to dispositional change by mere sentiment and thus alienating themselves from the pure practical reason that alone can supply the right sort of moral motivation.

My response to this objection has the following structure: either the objection depends upon a misinterpretation of the rationalizability requirement or it does not. If it does rely on a misinterpretation then it fails and if it does not rely on a misinterpretation then it must rely on a 'de dicto' interpretation of Kantian moral motivation or fail due to a resulting assimilation of the Kantian view to the rationalizability requirement. If the objection relies on a de dicto interpretation of Kantian moral motivation, however, then it fails for general reasons. Thus the objection must fail.

To address the misinterpretation first; what's demanded by the rationalizability requirement is the representation or presentation of relevant facts to the disposition-changing agent and the causal attributability of dispositional changes to this representation or presentation. This requirement makes no stipulations about the content of the causal chain that exists within an agent's psychology connecting these representations or presentations to the dispositional changes they inspire. This chain may involve sentiments such as sympathies or fears, or indeed beliefs (as in the subset of moral-dispositional changes that result from moral reasoning) or perhaps even neurons (if you prefer talking about such things), but these causal mediators are incidental. All that matters, as far

---

[337] H.Allison (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press; p.111

[338] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; p.38

as the rationalizability requirement is concerned, is that these causal mediators do their causal mediation properly; that they motivate dispositional changes relevant to the facts that trigger them. This makes it problematic to convict the rationalizability requirement of sentimentalism insofar as the role it gives to sentiments is limited to that of accidental causal mediation. It accepts that sometimes sentiments may trigger appropriate reactions to reasons, at least in some circumstances, and it tolerates such triggering whenever – and only whenever – things go this way.

The misinterpretation at work here is to read the rationalizability requirement's acceptance of sentiments such as sympathy or fear as mediators between facts and dispositions as a general defence of the role of sentiment in properly shaping moral dispositions. It is no such defence: it merely constitutes the acceptance of sentiments as mediators only when they happen to be good mediators – when they connect the right facts with the right dispositional changes (as they sometimes do). The rationalizability requirement doesn't make any claims, however, about just how good sentiments actually and generally are at doing this mediating work. They might be very good at it, very poor at it (as Kant worried[340]), systematically better or worse at it than other mediators and so on. All I suggest is that there's no good reason to assume sentiments generally unable to do this mediating work at the outset.

For this same reason the rationalizability requirement doesn't commit to a contingency of moral dispositions upon experience. It accepts experience as a possible route to good moral dispositions, when things work properly, but doesn't determine the right moral dispositions in terms of the outcome of some ideal set of experiences as per Hume[341]. The rationalizability requirement doesn't, indeed, commit to any claims about what the right moral dispositions are.

This misinterpretation aside, though, there remains a serious objection here. The objector may grant that the rationalizability requirement isn't implausibly sentimentalist, but still suggest that it errs insofar as it accepts moral-dispositional changes which result from psychological processes that don't involve proper moral motivation. You cannot merely, as the rationalizability requirement suggests, recognise the suffering of another (say), and thereby rightly be unproblematically moved to help suffering others. Rather *you must recognise the rightness of being disposed to help suffering others and work from that to the disposition to help.* You must establish good dispositions *under the description of doing what is right* if you are to achieve ideal agenthood (in the form of the Kantian virtue of 'autocracy')[342]. Only in doing this can you achieve the behaviour stemming from a good will celebrated by the Kantian school as holding moral worth[343].

There's some plausibility to this objection. We prefer well-intentioned failures to malicious successes. We place some value on the intent to do what's right, whatever it is. The rationalizability requirement seems to set such values aside in making its demands. It doesn't demand that we discern the facts, use them to figure out what are the right dispositions, and then adopt these right dispositions. Rather it merely asks that reasons themselves properly determine our moral dispositions, somehow. This misses out the step of recognising what is right that the Kantian

---

[339] *see ibid.* pp.92-115

[340] *ibid*. pp.37-38

[341] D.Jensen (2012), 'Kant and a Problem of Motivation', *The Journal of Value Inquiry*, vol.46, no.1, pp.83-86; pp.86-87

[342] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; p.51

[343] J.Timmermann (2009), *Kant's Groundwork of the Metaphysics of Morals*, New York, Cambridge University Press; p.49 [footnote]

plausibly requires, a step that seems worth valorising in a story of how we ought to go about changing moral dispositions.

It is my view that this objection depends upon an ambiguity in how we should interpret the Kantian doctrine. If the doctrine holds that we should form dispositions to act for the right reasons, then we may ask what makes these reasons the right reasons to form dispositions to act. Is it their nature – they are the right reasons because that's the way they are, in virtue of their properties – or is it our attribution of righteousness to them?

If we ought to develop dispositions in response to reasons in virtue of recognising them as being the right ones to respond to in this way, then this amounts to a particular thesis about Kantian moral motivation. This thesis analyses the defining content of this motivation in de dicto terms: it suggests that the important content of this motivation is the rightness accorded to it, not anything in its nature – whatever it may be – that goes beyond this ascription.

If, on the other hand, we ought to develop dispositions in response to the proper reasons out there in virtue of their nature (their actual properties) then this amounts to a different thesis about Kantian moral motivation. Per this thesis the defining content of this moral motivation is to be analysed in de re terms, as dependent upon the nature of the reasons (the facts) which properly inspire us to change our ways. What matters isn't that we recognise that a motivation is the motivation to do what's right but that we're moved by the right things under whatever description, whichever attributions of qualities. Per this second thesis you can be (pro-tanto) acceptably motivated by reasons which make a dispositional change right or at least good without necessarily recognising that the change is right or good (consider Huckleberry Finn, who describes helping the runaway slave Jim as wrong yet is moved to do it[344]); per the first thesis this is impossible.

These two interpretations relate differently to the rationalizability requirement.

The de dicto interpretation supports an argument against the rationalizability requirement. The requirement, after all, requires no necessary attributions of righteousness[345] to changings of moral dispositions in delineating which methods of changing moral dispositions are acceptable. True, there may be such attributions, as components or as side-effects of causal chains between facts and dispositions, but such attributions are treated as irrelevant to the legitimacy of the influences. matters for rationalizability is only that dispositional changes are caused by perceived reasons and are such that they're counted in favour of by these reasons, not that the subject of dispositional change attributes certain qualities to these dispositional changes along the way. As the rationalizability requirement doesn't require attributions of righteousness or other evaluative qualities but Kantian moral motivation, interpreted de dicto, does require such attributions, the two positions turn out to be inconsistent in theory and practice.

Consider that a campaign combatting excessive drinking, which confronts its audience with ugly drunkenness (say by vivid portrayals of drunken violence, of a sort whose genuine possibility I'll assume constitutes a good reason not to drink excessively), could be problematic for the Kantian autonomist thinking in de dicto terms. The campaign could be problematic insofar as it may condition behaviour through audience responses to the graphic drunkenness it depicts (some public information campaigns are supposed to work this way[346]) without at any point leading the subject of

---

[344] J.Bennett (1974), 'The Conscience of Huckleberry Finn', *Philosophy*, vol.49, no.188, pp.123-134; pp.125-127

[345] or any other evaluative quality one might propose

[346] L.Wallis (2013), 'Scared Smokeless: Graphic Antismoking Ads Increase Quitting Attempts', *The American Journal of Nursing*, vol.13, no.2, p.16; p.16

its influence to attribute righteousness (or indeed goodness or any evaluative quality) to not getting excessively drunk. The supporter of the rationalizability requirement, by contrast, instead asks of such a campaign that it work only by representing things which are good reasons not to drink excessively (for example, the carrying-out of drunken violence) and that it accomplish this representation successfully (in the way suggested in §3.10 and the more so the better). Such a supporter could thus endorse this sort of campaign, even if it works by repelling us from the ugliness of being excessively drunk, provided that being excessively drunk *is actually that ugly* and this is a reason to avoid such a state. In cases like this Kantian autonomism, trading on a de dicto analysis of moral motivation, clearly rejects the judgements implied by the rationalizability requirement.

However, there is no such clear rejection on a de re interpretation of Kantian moral motivation. On this interpretation what matters is that we respond to reasons however we describe things by our attributions of qualities, and this is precisely the demand of the rationalizability requirement. According to both the requirement and the Kantian doctrine with moral motivation understood de re, the acceptability of moral-dispositional changes varies according to the nature of the reasons that inspire said changes. Thus, cases like that of the public information campaign come out as unproblematic for the Kantian insofar as the motivations they inculcate turn out to be precisely the moral motivations required for acceptably changing one's moral dispositions. This is the case because in such cases the inculcation is being done by facts which constitute reasons to change one's morals. Recall, on a de re interpretation of moral motivation, what matters is that one is motivated by reasons' content, somehow. The upshot of the Kantian demand for the proper moral motivation, understood de re, thus turns out to be a doctrine similar in form and practice to the one implied by the rationalizability requirement.

If this assimilation works then Kantian autonomism only poses a genuine challenge to the rationalizability requirement if it incorporates a de dicto interpretation of moral motivation. Elsewise it seemingly leads one to endorse the requirement.

In settling matters between the rationalizability requirement and Kantian autonomism (and hence a sizable part of the freedom tradition), therefore, we must settle matters between de re and de dicto interpretations of Kantian moral motivation. We must determine whether, for the autonomist, moral motivation is better understood in de dicto or de re terms.

Kant himself occasionally seems to support a de dicto analysis of proper moral motivation[347], and there's a live literature supporting the place of the motivation to do what's right – whatever one calls right – in ethics[348]. Nonetheless there's also much in the autonomist tradition most easily understood through a de re interpretation of proper moral motivation. Kant himself accepted people coming to act rightly under the description of 'becoming children of God' and through other religious metaphors and fables, holding these to represent his ethical theory in a manner palatable to the masses[349]. This acceptance makes sense if one takes a de re view of the moral motivation Kant celebrated, in which case such metaphors can give you the proper moral motivation insofar as they manage to represent the reasons that give this motivation content. It makes less sense if one takes a de dicto view, insofar as this view predicates the possession of proper moral motivation upon developing moral motivations in response to attributions of rightness. One needn't make such

[347] *see for example* I.Kant (1797[1991]) trans. M.Gregor (1991), *The Doctrine of Virtue*, Cambridge, Cambridge University Press; p.59
[348] *see for example* R.Aboodi (2017), 'One Thought too Few: Where De Dicto Moral Motivation is Necessary', *Ethical Theory and Moral Practice*, vol.20, no.2, pp.223-227
[349] *ibid*. pp.97-98

attributions in observing religious rules; one may instead merely be trying to be holy (whatever this is) or follow what one takes to be a divine will.

As my goal here isn't exegetical, I won't make claims about how Kant or Kantians should be interpreted. I'll instead speak directly to the question of whether de dicto moral motivation seems like a thing particularly worth having. If it isn't, then it turns out that the version of the Kantian doctrine that cannot be assimilated to the rationalizability requirement fails – it fails insofar as the evaluative position which prevents its assimilation itself fails. This being the case there seems to be reason for the Kantian to support the rationalizability requirement.

The conventional complaint about a de dicto motivation to do what's right in ethics, owed to Michael Smith, holds that being motivated to act by one's attribution of rightness to an action rather than by the right-making features of said action involves 'fetishizing' morality. This fetishization consists in being motivated by attributions of moral qualities to actions rather than the contents that should motivate these attributions and actions[350]. This motivation is called a 'fetish' because it involves the misdirection of motivation away from its proper focus; from *the valuable* to *our attributions of value*[351]. Plausibly what really matters is what *is valuable* whether-or-not we call it such.

This 'moral fetishism' complaint may be recruited in opposition to the valorisation of Kantian moral motivation interpreted de dicto (which is different from the Humean strong externalism criticised by Smith in that it needn't invoke a desire to do what's right[352]; it involves motivation caused by attributions of righteousness but needn't specify the psychological states that do the motivating). Insofar as a de dicto interpretation of Kantian moral motivation essentially involves motivation to do what's right irrespective of the contents of 'doing what's right', this form of motivation seems to involve the sort of misdirected motivation that is sufficient for fetishism. It seems to do so insofar as it involves a fetishistic motivation by attributions of rightness rather than the right-making features – the bits of reality that are reasons – of Kantian ethics.

This moral fetishism imperils any Kantian account that prizes this species of motivation, whether it sees such motivation as sufficient to legitimate influencings involving it or only necessary for such legitimation[353]. This peril stems from the inherent unappealingness of moral fetishism, which is defined as involving misdirected motivations and a defective moral psychology. We, as aspiring moral agents, don't want our moral motivations to be misdirected or to have defective moral psychologies. Given that we don't want these things we shouldn't valorise de dicto moral motivations. Rather we should valorise motivation de re (in moral matters) and hence the sort of influencing endorsed by the rationalizability requirement.

This argument is dependent upon the charge of moral fetishism sticking to the proponent of de dicto Kantian moral motivation. This is a problem for the argument insofar as it leaves the argument open to an accusation of question-begging. This is because the fetishism charge will stick to the proponent of de dicto Kantian moral motivation only insofar as their account valorises misdirected motivation. The proponent of de dicto Kantian moral motivation's point, however, is precisely that to be

---

[350] M.Smith (1994), *The Moral Problem*, Malden, Blackwell; pp.74-76

[351] *ibid.* p.76

[352] *ibid.* p.74

[353] As in the case where the Kantian demands both de re and de dicto moral motivation; that the agent be moved by the right- or good- making contents of a reason *and* attributions of rightness or goodness to changes it encourages.

motivated by attributions of righteousness is a way – indeed *the way*[354] – to be properly motivated in moral matters. If this is the case then there's no reason to charge the proponent of de dicto Kantian moral motivation with moral fetishism, but to deny this is to seemingly deny the core of the proponent's account.

To evade this accusation of question-begging, it's necessary to interrogate the reasoning behind the proponent of de dicto Kantian moral motivation's valorisation of their favoured form of motivation. If this reasoning is flawed, the accusation of moral fetishism may be made to stick.

One reason to care about doing what is right, irrespective of the content of 'what's right', is that being motivated in such a way acts as a defence against problematic temptations. When one is motivated only by attributions of rightness one can focus one's mind entirely on the abstraction that is an attribution of rightness rather than needing to direct one's attention to the messy thing that is the world to find motivation. In doing so one can avoid the distractions and temptations of the world that may otherwise lead you away from your goals[355]. This sort of thought might lead one to the view that to care about doing what's right, de dicto, need not be fetishistic. It can rather be a motivation directed properly in accord with one's goals regarding one's behaviour[356].

The conventional response to this kind of reasoning is to worry that its focus on the valued in preference to the valuable allows for dangerously single-minded agents, who suppress qualms which otherwise help people to act rightly in contexts where their moral judgements fail them[357] (somebody who is motivated only by attributions of righteousness is incapable of inverse akrasia, which otherwise might help to moderate the consequences of bad moral judgement[358]). This response depends on there being inconsistencies between human attributions of righteousness and what is right, inconsistencies which result in more trouble when one makes oneself such that one is motivated only by attributions of righteousness. A defender of de dicto motivation could counter this argument by claiming that there are few such inconsistencies; that our attributions of righteousness, generally, are good guides to what's right.

The prevalence of such inconsistencies as well the actual effectiveness of de dicto motivation as a bulwark against temptation are matters for debate and likely vary person-to-person. Some may be destructively akratic such that they would act better if strongly moved by their ascriptions of value; some may regularly miss-ascribe value and act better if moved by what's valuable rather than ascriptions of value. There's little I can add to this debate, save a worry that it will require significant commitments to settle. To settle it one would have to work out what we ought to do in all circumstances and then determine whether de dicto or de re motivation (or some workable carve-up between the two) would, overall, get us closest to doing what we ought to do all the time. As I do not mean to presume any comprehensive first order ethics nor have the space for such determinations, I cannot settle such matters here.

Even where no inconsistencies between attributions of righteousness and what's right occur, though, in my view there's still a problem with the character who reconstructs their motivations in

---

[354] I.Kant (1797[1991]) trans. M.Gregor (1991), *The Doctrine of Virtue*, Cambridge, Cambridge University Press; p.59

[355] J.Olson (2002), 'Are Desires De Dicto Fetishistic?', *Inquiry*, vol.45, no.1, pp.89-96; p.92

[356] H.Lillehammer (1997), 'Smith on Moral Fetishism', *Analysis*, vol.57, no.3, pp.187-195; pp.191-193

[357] V.Carbonell (2013), 'De Dicto Desires and Morality as Fetish', *Philosophical Studies*, vol.163, no.2, pp.459-477; p.462

[358] C.Kleist (2009), 'Huck Finn the Inverse Akratic: Empathy and Justice', *Ethical Theory and Moral Practice*, vol.12, no.3, pp.257-266; pp.264-265

keeping with a requirement to be motivated only by attributions of righteousness. This problem is – for want of a better word – a problem of idolatry; a problem of treating a representation of a thing as the thing itself. This problem remains irrespective of whether those motivated by their attributions of righteousness or goodness more reliably do what's right or good.

The traditional problem with the idolater is that they mistake the more-or-less representative parts of the faith they aspire to maintain for parts of its content. The idolater regards statues of their gods as their gods or treats their books as holy rather than the things they communicate, say. In doing such things it's tempting to say that the idolater goes wrong.

The reason for this is that representations (somehow) get their content from their relations to the things they represent. The status of a map as a representation of an area depends upon it standing in certain relations to the area it represents, such as structural correspondence or proper genealogical connection (being made by a cartographer based on observations of the area, say). The status of a word as a representation of something also depends upon some such connections (perhaps established by a naming event, or a conventional linguistic role). When a representation is employed without an awareness of these kinds of connections, and their role in making the representation meaningful, the representation is misunderstood. One becomes unable to recognise the important part of what is shared between the words 'white' and 'blanc' and tells the francophone he is wrong in claiming that the colour of snow is 'blanc' for it is actually 'white'. One becomes rather foolish in the matter.

The analogous mistake of the idolater is to forget that whatever meaning representations have they have only in virtue of relations with the things they represent. Instead, the idolater makes his representations into idols; things possessed of inherent meanings derived from the different things they once represented. On the practical side the result is much silliness; debates over who has the 'holiest' trinkets supplanting sensible theology. On another side emphasised by those who stress the idea of a connection to the divine in religion the result is a troubling alienation.

Analogously, the mistake of the apologist for de dicto Kantian moral motivation is to forget that the meaning of the representations that constitute their beliefs attributing rightness to given moral dispositions depends upon reasons, as found in the messy world they suggest we shouldn't be motivated by (at least in some significant part). In valorising de dicto Kantian moral motivation the apologist for it ignores this fact and thereby makes idols of his representations of what is right. He cuts these representations loose from the reasons that give them meaning and invests them, confusedly, with the same meaning as if nothing has changed.

The result is a distinctive error; the value of acting in accord with one's beliefs supplanting the value of doing what is right or good. This results in a valorisation of merely continent agents for whom values are supplanted by the injunction to do what they believe to be right[359] (consider, here, Anscombe's unsympathetic rendering of the Kantian moral agent as too trapped within description-relative evaluations to properly respond to the valuable itself[360]). A further analogous result is an alienation from the reasons the rationalizability requirement suggests our moral-dispositional changes should be connected with.

The charge of moral fetishism thus turns out to stick after all. Being motivated to do what's right de dicto is to fetishize doing what's right inasmuch as this involves having a misdirected motivation; in

[359] A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; pp.40-41

[360] I.Schumski (2017), 'The Problem of Relevant Descriptions and the Scope of Moral principles', *European Journal of Philosophy*, vol.25, no.4, pp.1588-1613; pp.1588-1589

this case, a motivation directed towards a representation rather than what it represents (or should represent).

If this charge sticks it seems we ought not valorise or accept de dicto Kantian moral motivations in forming judgements about what is and is not acceptable in moral-dispositional change. To do so would be to make idols out of one's moral beliefs and end up not only courting error but alienated from precisely the moral content that one's moral dispositions ought to have. If this is the case, then there can be no challenge that may be mounted to the rationalizability requirement from a position that supports the place of de dicto Kantian moral motivation in moral-dispositional changes for the general reason that no such position is defensible. If one is a Kantian autonomist, thus, one should valorise only de re moral motivations, and thus – if I'm right – support the rationalizability requirement as a way of delineating legitimate influences.

(5.6) Beyond Autonomy: Interpersonal Freedom

I've argued that only representations or presentations of relevant truths – reasons – can legitimately cause moral-dispositional changes in an important pro-tanto sense, and that this view coheres with an autonomist view of freedom which identifies freedom with acting under the compunction of de re reasons. Autonomist views, though, don't exhaust the concept of freedom as it might impinge upon debates about which influences on our behaviour we should tolerate. You might value freedom in some other sense.

Particularly, you might value freedom as something got by being in certain ways independent of other agents. You might think freedom is about not being coerced or dominated by anybody. You might care about securing something like what Berlin popularised as 'negative liberty', a liberation from the interference of other agents[361]. You might demand Pettit's 'republican freedom' which requires liberation from interference by others who aren't compelled to consider your interests[362]. You might, generally, understand freedom as there not being some other agent with inappropriate power over you, including over your dispositions. If you want freedom in these or relevantly similar senses then you want a freedom which is necessarily *interpersonal* in a way that freedom isn't for the autonomist, for the autonomist conception of freedom is ultimately directed at the goal of rational action (it's 'positive' in Berlin's sense[363]) and this goal may be frustrated by your own peculiarities or the machinations of impersonal chance as well as the actions of others. These other sorts of freedom, by contrast, idealise certain isolations from other agents – to be free of all the interventions of others or at least certain domineering ones[364]. As such these other sorts of freedom may only be threatened by the actions of other agents.

Arguably, for these interpersonal freedoms, amongst the actions deleterious to your freedom may sometimes be showings of relevant reasons of the sort that my theory legitimates as acceptable means by which to change moral dispositions.

An honest propagandist (perhaps such creatures exist) might use her propagandising skills to show you an accurate representation of a bitter truth, relevant to your behaviour, and thereby change it against your will. You might enjoy eating cakes but lose your desire to eat them once you're presented with facts about what they'll do to your health. You might like using a drug but be so

---

[361] A.S.Kaufman (1962), 'Professor Berlin on 'Negative Freedom'', *Mind*, vol.71, no.2, pp.241-243; p.241
[362] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; p.55
[363] K.Möller (2012), *The Global Model of Constitutional Rights*, Oxford, Oxford University Press; pp.29-30
[364] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; pp.65-67

shocked by relevant facts about the brutality of its production that you stop using it. You might prefer a certain political candidate but find yourself unable to continue supporting them following exposure to relevant facts about the effects of their policies. You might lament each of these changes and see them as interferences in your behaviour which compromise your freedom to eat what you want to, use the drugs you want to and support the politicians you want to. You might sing this lament despite acknowledging the truth and relevance of the representations supplied by the honest propagandist, precisely because you think freedom consists in making your own choices free from others' interference no matter how this interference works. When your complaint has this form it cannot be stated in terms of autonomy and must invoke some 'interpersonal' conception of freedom.

One option I have in responding to such complaints is to set them aside. This thesis is meant to describe values distinctive to processes of influencing behaviour; it isn't a comprehensive treatment of everything we value that may impinge upon such processes. If we value interpersonal freedom in the way suggested above I may treat this valuing as something to weigh in practical judgement against the value of being inspired only by relevant facts. I could avail myself of this option with respect to interpersonal freedoms, but I will not do so (although I do invoke this defence with respect to non-autonomist goal directed 'freedoms' which identify freedom with things like 'promoting the creation of a given sort of society[365]'). I will not do so for two reasons.

The first is the central place of invocations of interpersonal freedom in thought about the ethics of influencing moral dispositions. A significant minority oppose influences perceived as controlled by others seemingly whatever the mechanisms or goals of said influences[366]. This stance seems to only be sensible in terms of commitment to interpersonal freedom. Very – perhaps *most* – often, when concerns are raised about methods of dispositional influencing they're raised in something like the following way: 'this method gives those who use it too much power over others![367]' To set interpersonal freedom aside would thus be to side-line a particularly important part of what people consider important when they make the decisions I mean to counsel.

My second reason for not setting interpersonal freedom aside is that I believe there are good reasons for thinking it is properly preserved by influencing compliant with the recommendations of the rationalizability requirement. I think concerns about honest propagandists are misplaced and not because I think that applying the rationalizability requirement is generally more important than promoting interpersonal freedom. I think rather that promoting interpersonal freedom in relevant contexts encourages applying the rationalizability requirement.

I think this because, frankly, one cannot tyrannise by truth-telling. You cannot gain power over another person by simply showing him the truth, at least if this is all you do. You can – in a sense – give the truth itself power over someone by showing it, but the truth is not an agent with the capacity to threaten interpersonal freedom nor can it come under the control of an influencer such that its power is passed to the influencer.

---

[365] *see for example* L.P.De La Escosura (2015), 'Human Development as Positive Freedom: Latin America in Historical Perspective', *Journal of Human Development and Capabilities*, vol.16, no.3, pp.342-373; pp.342-343

[366] W.Hagman, D.Andersson, D.Västfjäll and G.Tinghög (2015), 'Public Views on Policies Involving Nudges', *Review of Philosophy and Psychology*, vol.6, no.3, pp.439-453; pp.451-452

[367] C.Bublitz (2016), 'Moral Enhancement and Mental Freedom', *Journal of Applied Philosophy*, vol.33, no.1, pp.88-104; p.103, *see for example* J.Harris (2014), 'Taking Liberties with Free Fall', *Journal of Medical Ethics*, vol.60, no.6, pp.371-374; p.372

This all being the case there will never be a circumstance where influencing compliant with the rationalizability requirement compromises interpersonal freedom. This being the case anyone who values interpersonal freedom, at least in influencing contexts, has reason to accept the rationalizability requirement.

Before arguing these points, I'll first detail the kind of interpersonal freedoms that I'll be addressing and show why one might initially think them in tension with the rationalizability requirement.

(5.7) The Challenge of Interpersonal Freedom

I've mentioned Berlin's 'negative freedom[368]' and Pettit's 'republican freedom[369]' as touchstones for the kind of interpersonal freedoms I mean to discuss. They're different conceptions but it's what they have in common that I mean to speak to here.

Berlin's 'negative freedom' identifies being free with the absence of external agents interfering with you – making you do things, stopping you from doing things, and meddling with the outcomes of your actions[370]. It's thus a concept which sees sacrifices of freedom inextricably involved with a great many ordinary interpersonal activities such as tax-paying, indeed anything that impacts your choices in a non-trivial way and proceeds without your consent[371]. The 'republican freedom' defined by Pettit originated in a reaction against this conception of freedom, or at least the dichotomy Berlin proposed between it and goal-directed 'positive freedoms[372]'. The concept still demands the absence of interference by external agents but grants that such interference may not compromise an agent's freedom by creating interpersonal power relations provided that said interference is compelled by some appropriate mechanism to track the interests of the agent being interfered with. Hence being taxed, say, needn't disrupt freedom provided the tax-raising authority is compelled to consider one's interests[373].

What such accounts (which I'll argue stand in no direct tension with the rationalizability requirement, though one might argue otherwise) have in common is the claim that what compromises freedom, of a kind we may value, is interference with the free agent by external agency. With republican freedom interfering agents have access to a procedural get-out clause; with negative freedom this escape is unavailable (and consent is required for freedom-preserving intervention). Either way, though, the way you destroy someone's freedom is by meddling with the choices they make; by making them make certain choices rather than others, or by imposing penalties on certain choices, and so on. This destruction, in turn, can only be accomplished by an external agent; one's freedom can't be compromised by nobody or oneself.

Consider §5.6's 'honest propagandist'. The honest propagandist has an agenda for your behaviour; she wants it to be a certain way. She gets what she wants by showing you certain reasons which she knows (for she's skilled at her trade) will compel you to change your behaviour in the way that she wants it changed.

---

[368] I.Berlin (2002 [1957]), 'Two Concepts of Liberty' in H.Hardy [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217; pp.169-170

[369] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; pp.66-68

[370] A.S.Kaufman (1962), 'Professor Berlin on 'Negative Freedom'', *Mind*, vol.71, no.2, pp.241-243; p.241

[371] I.Berlin (2002 [1957]), 'Two Concepts of Liberty' in H.Hardy [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217; p.170

[372] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; p.18

[373] *ibid.* pp.63-64

It's prima facie intuitive to see a sort of power in such an interaction, such an influencing. *That* the thing this power was over seems to be your dispositions and *that* the power was exercised by an agent other than you seem to make this power threaten your interpersonal freedom. Freedom, in all interpersonal senses, demands that there never be such power relations. It demands that nobody else ever control your behaviour or at least that they only do so with your consent. The honest propagandist, let's say, neither has your consent when she effects change in your moral dispositions (she, say, simply plasters a representation of certain compelling facts on a billboard and leaves it somewhere where she knows you'll see it) nor is she in any way compelled to consider your interests in making decisions about influencing you (which one might think could protect your freedom[374]). As such the honest propagandist seems to jeopardise your freedom. She deprives you of the ability to choose your own moral dispositions and makes the choice for you by employing the power she has, in virtue of what she knows and can do, over your moral dispositions. She may do this despite your wishes and complaints. She might perhaps even find some way to force the relevant truth into your view, even as you try to ignore it.

If such influencing can indeed compromise an agent's interpersonal freedom it follows that this freedom is undefended by my rationalizability requirement. The requirement, after all, cannot restrain the honest propagandist. Provided such a propagandist influences only through showing relevant reasons – as they do, ex hypothesi – per the requirement they're doing nothing objectionable. Indeed, what would be objectionable, per the rationalizability requirement, would seem to be precisely the steps one might recommend to protect interpersonal freedom in such cases. One might, for example, recommend averting one's attention from the relevant reasons the honest propagandist shows to save your interpersonal freedom from their machinations.

Nonetheless, seemingly in defence of interpersonal freedom and defiance of the rationalizability requirement, some propose we should act in such ways. Think for instance of the complaints sometimes raised against public health campaigns which decry them in a language of interpersonal freedom, castigating the state for its paternalistic interference in the otherwise free choices of citizens[375]. Such complaints needn't dispute the reasons or the representations of reasons within such campaigns; they may instead dispute whether these campaigns can justly combine the propagandising resources of the state with these reasons to shape public behaviour[376]. A properly free public shapes its own behaviour, reasons notwithstanding, so goes the worry[377]. Analogous complaints are sometimes formulated regarding leaked information. It is worried that the information thus revealed, though not necessarily false or irrelevant, has the potential to interfere with otherwise free behaviour[378]. So-called 'post-truth' politics[379] may also be understood as justified by the view that agents should forge their own positions free from the interference of others, even others telling relevant truths. In this way of politicking 'truth' may be viewed post-structurally as

---

[374] *ibid.* pp.63-64

[375] *see* G.Tocker (2011), 'The Risk of the Nanny State: Lifestyle Advice in Public Health Campaigns', *Arbeiten aus Anglistik und Amerikanstik*, vol.36, no.1, pp.3-16; pp.8-9, 14

[376] *ibid* pp. 3-4, 14

[377] *see for evidence of prevalence* M.L.Crossley (2002), 'Resistance to health promotion: a preliminary comparative investigation of British and Australian students', *Health Education*, vol.102, no.6, pp.289-299; p.295

[378] *see for example* BBC News (2019), 'General Election 2019: Source of UK-US trade document leak must be found – PM', *BBC*, Available at https://www.bbc.co.uk/news/uk-50699168 [webpage] (accessed 12/12/2019)

[379] *see for example* G.Hewitt (2017), 'Trump and truth', *BBC*, Available at https://www.bbc.co.uk/news/world-us-canada-38731191 [webpage] (accessed 10/02/2020)

merely rhetorical (in the manner popularised by Foucault[380]) or else as simply less important than extant feelings and preferences. Either way the position denies that the truth, or some special set of truths (reasons), should hold some venerated place in debate; its influence should be challenged as a threat to people's ability to think freely and as a tool of power.

I believe – contra such concerns – that implementing the rationalizability requirement in managing influences on at least one's moral dispositions can help protect one's interpersonal freedom. I believe this because I believe that *there will never be any direct practical disagreement between the maximisation of interpersonal freedom and discrimination amongst influences according to the rationalizability requirement*. There may perhaps be indirect disagreements as in cases, say, where one is influenced in ways which somehow indirectly risk one's interpersonal freedom (suppose one is shown a reason to become more compliant and hence made such that one is in the future more vulnerable to domination). In direct cases, however, where another's influence appears to at once be rationalizable and to threaten one's interpersonal freedom (as with the influence of the honest propagandist) this appearance will always be illusory. If such appearances are always illusory then interpersonal freedom cannot be directly threatened by any method of influencing moral dispositions which meets the rationalizability requirement. If interpersonal freedom cannot be directly threatened by any method of influencing moral dispositions which meets the rationalizability requirement then the valorisation of the former doesn't generate any general challenge to the latter. On the contrary; it suggests the latter may be used to aid in the protection of the former by helping to discriminate between influences which do and do not threaten interpersonal freedom.

(5.8) Asymmetric Means

One concern I'd like to dismiss immediately – which might be thought to support the view that influencing by showing relevant reasons can compromise interpersonal freedom – is the concern that some agents are more able to disseminate representations of relevant reasons or render them impactful (through vividness) than others. This can result from all sorts of asymmetries; some might have access to greater skill at reason-showing, more money to hire those with such skills, or perhaps just more or more compelling reasons with which to work. Entities like powerful governments and businesses can employ greater resources to influence those they want to act differently than the average agent. One might suggest that the resulting asymmetry breeds dysfunctional games of mutual influence and these worries are best analysed in terms of interpersonal freedom. Plausibly what makes games of mutual influence in which one player starts with greater resources bad is the position of power this puts the lucky player in with respect to other players. The influence involved stops being fully reciprocal, and some find this problematic[381].

Such asymmetries strengthen worries about threats to interpersonal freedom less than it might seem at first.

Imagine a world unlike our own in which influencing resources are evenly distributed, a world in which no agent can call upon more money, greater skill at influencing or even more or more compelling reasons with which to influence her neighbour than her neighbour can call upon in influencing her. Imagine further that in this imagined world aside from this difference people live as

[380] S.Prozorov (2019), 'Why is there Truth? Foucault in an Age of Post-truth Politics', *Constellations*, vol.26, no.1, pp.18-30; p.18

[381] J.Lenman (2014), 'Gibbardian Fallibility: Moral Fallibility and Moral Smugness', *Journal of Value Inquiry*, vol.48, no.2, pp.235-245; pp.240-241

they do in our world, constantly using whatever resources they're willing to employ to influence others to act in ways that they want them to act.

Such a world would seem to be far from devoid of the sorts of power decried by proponents of interpersonal freedom. You'd still live at risk of having your behaviour determined by the influence of outside agents. The game of influence you play with them would be *fair*, in a gamblers' sense (nobody would start with an advantage), but it would still be a game of influence and with respect to the value of interpersonal freedom this is precisely the problem. To take the value of interpersonal freedom seriously is to think that nobody else should be able to determine the way you act; it matters nothing whether you are equally able to determine the way they act. Two wrongs don't make a right. Retribution notwithstanding, my attacks on your freedom don't alter the evaluative qualities of your attacks on my freedom.

In further discussing interpersonal freedom in the context of the ethics of moral-dispositional change I'll thus set such asymmetry concerns aside. It's certainly true that in general greater influential powers pose greater risks to interpersonal freedom but the problem here is the magnitude of such powers and the consequent risks they pose, not their unequal distribution. It may also be true that in a better, fairer world some or all influential capacities would be more evenly distributed but it isn't clear how such more even distributions could negate risks to interpersonal freedom arising from the use of such capacities[382].

(5.9) Killing Messengers and the Influence of Reason-showing

I've proposed that there will never be any direct practical disagreement between the maximisation of interpersonal freedom and discrimination amongst influences according to the rationalizability requirement. I believe this because I believe that the reason-shower can't exert influence.

They might mediate the influence of relevant reasons and in doing this form part of a causal chain which determines which influences agents will be subjected to and do so precisely to see certain agents influenced in certain (perhaps predictable) ways. Yet, I argue, none of this is sufficient to turn a reason-showing agent into a source of influence responsible for – and in possession of interpersonal power over – shifts in the dispositions of another. In showing relevant truths (as the rationalizability requirement requires) what you do is mediate the influence of a fact – a reason – which is such as to take (and not share) responsibility, in a certain sense, for any influence it exerts over behaviour. To be influenced by a reason-showing is to be influenced by a reason and not a reason-shower.

Given this, it turns out that there's no objection that the proponent of interpersonal freedom can make to any influencing by showing reasons. This is the case because to be influenced by being shown a reason is to be influenced by a reason and a reason is not a person. Only the influence of persons can threaten interpersonal freedom. Thus, no reason-shower has ever really exerted power over another agent by their reason-showing and no reason-shower has ever, by their reason-showing, directly compromised anybody's interpersonal freedom. Changes in behaviour triggered by

[382] It may be true that where influential capacities are more evenly distributed there will be tighter limits on the damage to freedom the exercising of these capacities can do, since people will be more able to influence others towards not influencing them in ways destructive of their freedom. Such limits, if they exist, only do so as contingent functions of how people use their influence over others and thus do not make more even distributions of influential capacities necessarily less hazardous to interpersonal freedom. If I fail (for whatever reason) to manipulate you into not manipulating me my capacity to manipulate you fails to protect my freedom from disruption by your manipulations.

the reasons displayed by such reason-showers are the responsibility of the reasons they show, not them as their mediators.

As such anybody who takes interpersonal freedom seriously will have reason to support the application of the rationalizability requirement to the evaluation of influences, at least on moral dispositions. They will have this reason because the rigorous application of this requirement to such influences should be able to guarantee they never directly threaten the interpersonal freedom of those subjected to them.

This argument requires support and explanation. The seemingly counterintuitive claim it depends upon is the claim that, somehow, to influence with a relevant truth is not to influence at all (though of course it may be to do other things – such as scare – which may be good or bad in various ways). This position seems instantly strange, indeed inconsistent with my previous commitment to describing such influencing as a 'method of moral-dispositional change'. Methods are used by agents, yes? Surely their results are agents' responsibility? As such it seems inconsistent (also *odd*) for me to claim that to exert influence by showing a relevant truth isn't to exert influence at all.

Given these apparent problems with my argument, I'll first justify my claim that when one influences by presenting or representing a reason one does not become responsible for the effects of this reason upon dispositions. In §5.10 I'll then show how this analysis is consistent with understanding 'truth showing' (by which I mean just influencing compliant with the rationalizability requirement) as a method of moral-dispositional change, indeed a method usable by agents.

It isn't novel to claim that the reason-shower shouldn't be held responsible for the effects their reason showing has on agents' dispositions; the view has antecedents in conventional understanding.

Some journalists, say, deny responsibility for changes in behaviour caused by the stories they tell. They – confident in the relevance and truth of their stories – say that they simply reveal what is the case, show things 'the way they are', and leave it up to this truth to have whatever effects it will on people's behaviour[383]. In doing so a journalist, despite being part of a causal chain ending in changed dispositions, disclaims responsibility for the outcomes of this causal chain. He does so because he thinks that the causal link which connects his actions to the changes in question is not such as to make him responsible for those changes. He doesn't think himself in control of his audience. The causal link which connects the journalist's actions to the behaviour of his audience – if the journalist has done his job right – is composed of relevant facts (reasons) vividly shown, exactly those things which I suggest aren't responsibility-preserving.

This way of denying responsibility isn't only employed by journalists. When we're attacked for revealing reasons, on account of the impact these reasons have on behaviour, we feel we're done an injustice. In such cases we understand ourselves not so much as agents imposing our will upon others, but rather as mediators for the impact of things beyond ourselves. We think it's wrong to 'kill the messenger' precisely because the messenger shouldn't be held accountable for his message – if it's relevant and true, anyway. The content of the message is not the messenger's responsibility[384].

---

[383] S.Tait (2011), 'Bearing Witness, Journalism and Moral Responsibility', *Media, Culture & Society*, vol.33, no.8, pp.1220-1235; p.1232

[384] This isn't to say that there aren't norms – of manners and broader consequence, say – dictating which messages you can acceptably deliver and when you can do so. Loudly telling someone they're fat in a public place can be wrong even if they are indeed fat and their fatness is somehow relevant. Tactfully pointing out to

How far, though, does this point extend? Responsibility isn't a monolithic thing but a matter of degrees and kinds. It makes sense to talk about someone as more or less responsible, sharing or monopolising responsibility. Relevantly, one might suggest that causal mediators insulated from the effects of their mediation by relevant facts, such as journalists, bear *some* or *some kind of* responsibility for the effects of their reason-showing. How then are we to explain and justify reason-showers' disclaiming of responsibility for the influence they seem to exert? It may be that this disclaiming is just a conventional way of talking with no deeper meaning, or simply wrong.

Suppose a journalist reveals that there's pollution in the public water supply and thereby causes people to avoid using the public water. Suppose further that this rationalizable avoidance ruins the water company and causes a disease outbreak due to reduced sanitation. There's a temptation in this case to call the journalist 'partially responsible' for the ruination of the water company and even the outbreak. There's even more temptation if we suppose that the journalist's report acted directly on the dispositions of the people who stopped using the water. Suppose for instance everybody already knew that the water was polluted but the report, without stooping to the use of influential non-reasons (like a sinister soundtrack) made the fact vivid in a compelling way and thereby motivated people to adjust their behaviour. Suppose for instance the report carried graphic pictures of poisoned fish floating in the visibly polluted water. Suppose further that the strong response to these pictures, and its effects on the water company and sanitation, were predicted by the journalist. How can the journalist not be at least 'partially responsible' in such a case, even though she only mediated relevant facts?

What this proposes is that, for the journalist, the relevant truth functioned at least partly as something like a *tool*; an instrument which passes responsibility on to those that use it to do things (such as tell stories about how things are). The journalist is – it is suggested – at least partially responsible for the company's ruination and the disease outbreak insofar as she used her tool recklessly, damaging the water company and public health. Doing so may have been required by her professional duties as a journalist, but nonetheless per this understanding she may be blamed (at least partly) for the trouble caused by her reporting (as Peter Stockman claimed in *An Enemy of the People*[385]). She may be blamed because the facts she reported were tools she used for influencing, and we're responsible for the effects of the tools we use. The journalist exerted influence over her audience by showing them relevant truths and is thus responsible for the consequent changes in their dispositions.

The journalist's professional ethics may protect her from being blamed for acting in this way, if they (say) demand that she report even troublesome things, but not in the way needed to make true the view that reasons don't pass along responsibility to those that show them. It may be that the professional ethics of journalism are out of tune with right and wrong proper (they permit troublesome reporting, broader ethics prohibits it) or are supported by other values (it may be good to have troublesome reporting so that, say, society is made aware of hidden problems). Either way, though, invoking these ethics as one's defence here doesn't contest the idea that the journalist has *some responsibility* for the effects of her report. It is to hold that the journalist did her duty as a journalist but she could nonetheless be to some extant responsible for the failure of the water company and the disease outbreak. This is the analysis I contest (at least where reports only influence by representing relevant facts).

---

someone that they're fat in the context of a private and frank conversation about such things, though, is another matter.

[385] H.Ibsen (1882[2018]) trans. R.F.Sharp (2018), *An Enemy of the People*, Urbana (Illinois), Project Gutenburg, available at https://www.gutenberg.org/files/2446/2446-h/2446-h.htm [accessed 14/1/2022]; Act II

The reason I contest this analysis is that I think that it's always and entirely between agents exposed to a reason and the reason itself to settle what dispositional changes follow from their exposure to said reason. Granted, any dispositional changes caused by such an exposure may be predictable, not chosen by agents in any meaningful sense or downright intended by the exposer of the relevant truth. They may also lead agents to act wrongly (as there can certainly be strong – though not sufficient – reasons for doing wrong). Nonetheless, though, I don't think that we should hold exposers of relevant truths responsible for the effects on moral dispositions that result from their truth-showings. I think this because I think – in cases where reasons are shown and reacted to properly (see §4.5) – we should (in a certain sense) blame the reasons shown, perhaps blame those who respond to these reasons and never blame the reason-shower for effects on dispositions (or causally downstream events).

Suppose, to return to the pollution example, you say the person who reveals the water is polluted is to some extent responsible for people refusing to use it and the resulting troubles. In attributing responsibility in this way, and thereby instrumentalising reasons (making it such that they at least partly pass responsibility on to those that use them in accomplishing their purposes, like tools) one redirects one's focus in evaluating influences from the world one inhabits, and the reasons with which it may move you, to the things that mediate that world to you in various ways. Doing so carries the risk that one will find oneself moved not to respond to reasons in one's actions, but rather to the people and mechanisms that show one reasons. You'll, at least to some extent, blame the person who reveals the water company's polluting for people's reactions to the pollution, and not the pollution itself or indeed the polluter.

In this way your behaviour will not be shaped by reasons so much as by a mix of reasons and facts about how you encounter them. How reasons come to one, I suggest, *shouldn't change how one responds to them* even though, as a matter of human frailty (and as discussed in §3.10), it may modulate their influence[386]. We should not evaluate the influence of reasons in a way that adds to this modulation by judging such influences by who mediates them or why (and other facts about said mediators) rather than just in terms of their contents[387]. In doing so one only grants oneself an excuse to ignore relevant reality in responding to influences upon one's behaviour. One creates scapegoats between oneself and reasons rather than letting them shape your dispositions as they ought to do.

This argument trades on the Nozickean intuition I introduced in §3.7[388]. It's important that one's behaviour is connected properly to reality, perhaps ultimately for remote practical reasons which structured human evolution and ultimately in justificatory terms because it's something we value. Allowing facts about who is showing you a reason and why they're doing it to interfere with its influence on your behaviour corrupts this valued connection. This is, however, precisely what one must do if one is to meaningfully attribute responsibility to reason-showers for the effects of their reason-showing. After all, if who shows you reasons and why they show you them play no proper

---

[386] P.Katopol (2018), 'The Halo Effect and Bounder Rationality – Limits on Decision Making', *Library Leadership and Management*, vol.32, is.3, pp.1-5; pp.1-2

[387] This doesn't mean that assessing the contents of represented reasons needn't involve considering facts about who represents them to us or why they do so. The contents of a represented reason plausibly depend on its truth – if such a reason lacks truth it will also lack content (at least, the content relevant to the rationalizability requirement) – and in assessing the truth of a representation it can be profitable to consider facts about whoever supplies the representation to us and their goals. Facts about influencers can evidence whether influential representations are reasons, but this isn't the same thing as these facts making it the case that reasons themselves are more or less legitimate influences.

[388] R.Nozick (1972), *Anarchy, State and Utopia*, New York, Basic Books; pp.42-45

part in your reaction to them, what could it mean to hold them responsible for your reaction? When we hold agents responsible for things, we direct our response to the thing we hold them responsible for towards them. To be 'responsible' is to be worthy of a certain sort of response or reaction – such as praise, blame or intervention[389]. When we respond properly to the influence of a relevant truth (not that we always do so), no part of this response varies with or targets who- or what- ever showed us said truth or any feature of them[390] – the response entirely varies with and targets the reason represented by the truth.

Return to the example of the journalist and the water company; who should we hold responsible for the company's failure and the disease outbreak? We shouldn't hold the journalist responsible. The journalist exposed reasons which exerted influence which led to the changes in behaviour that damaged the company and reduced sanitation, true, but in doing this the journalist was simply doing what she should have done. Not merely as a journalist, but as a human being in a particular context, the journalist found herself in a position to influence people and she did so, legitimately, by revealing relevant facts to them. If there's anything in this story to hold responsible it is not to the messenger we should look but *to the reason represented by the message*. We should look to how the pollution got in the water and hold those who put it there responsible (by blaming them) and we should also (in a sense) 'hold the truth responsible' by working to change it (make an intervention), concentrating our efforts on the reality of the situation rather than the irrelevancy that is who revealed it to us or why. We should get the pollution out of the water and maybe blame whoever put it there, not complain that our freedom was compromised when we were compellingly shown or reminded it was there. We might also have reason to blame those who were influenced by the journalist's report for the failure of the water company and other problems if, say, their rationalizable reactions to the journalist's report were somehow their choice in some meaningful sense (which they may or may not have been[391]).

The injunction not to hold the journalist responsible doesn't vary with the intentions of the journalist who showed us the pollution. The journalist may have been acting out of a desire to educate the public, to earn a salary or to benefit a competing water company run by her sister-in-law. She might have been influencing in good faith, showing reasons for what she believes to be a justified shift in dispositions, or bad faith, showing reasons for a shift in dispositions she believes is wrong for those influenced by her reporting. No intentions or states of influencers (alone) can ever be sufficient to make any difference in proper reactions to influencings themselves (as opposed to influencers themselves) and render influencers meaningfully more responsible for their influencings' effects.

This requires mention because one might (I think wrongly) regard facts about influencers and their intentions as non-evidentially important in deciding which influences to accept. Gideon Yaffe, for example, suggests 'manipulative' influences may be characterised by an influencer's intent to reduce the range of future options available to an agent[392]. Very generally manipulativeness in influences is often connected to certain intentions, such as the intent to deceive or more broadly to induce faulty

---

[389] M.Montminy (2018), 'Culpability and Irresponsibility', *Criminal Law and Philosophy*, no.12, pp.167-181; p.177

[390] Other parts of our responses to specific representations may vary with who- or what- ever originated the representation; such as when the representation (say) breaches a confidence.

[391] Depending, at least, on the truth of certain claims about volition; *see for example* C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; pp.164-166

[392] G.Yaffe (2003), 'Indoctrination, Coercion and Freedom of the Will', *Philosophy and Phenomenological Research*, vol.67, no. 2, pp.335-356; pp.343-344

mental states[393]. If influencers' intentions can render influences manipulative (say), and we ought to reject manipulative influences, it would follow that these intentions would make some difference to proper reactions to influencings.

Facts about influencers and their intentions can do things such as render influencers disingenuous or irresponsible. We, say, are likely to think poorly of an influencer who only shows a specific set of the relevant facts they know to ensure that the subject of their influencing is influenced as they intend them to be. We'll regard such an influencer – especially if the totality of their beliefs if represented would influence in a significantly different way (say against making the dispositional changes intended by the influencer, as in a bad faith case) – as disingenuous and plausibly blameworthy for this disingenuousness. Equally an influencer who is reckless or malign in considering the consequences of their influencing might still be accused of being irresponsible and blameworthy *for this* recklessness or malignity even if they influence only by showing reasons[394].

Such evaluation of influencers must however be distinguished from the evaluation of influences. Even a narrowly self-interested and partial reason-showing with manipulative intent is still a reason-showing and as such may acceptably exert influence on an agent's dispositions. The truth revealed by the disingenuous person remains true. The reason revealed by the would-be tyrant remains a reason. Evidence of disingenuousness or manipulative intent gives us reason to assess the truth and relevance of people's claims not because these things in-themselves imperil truth and relevance, or impact how we should react to reasons, but rather because the disingenuous or manipulative are presumed to care less about truth and relevance in choosing what to say and show. For intent to impact evaluation of influences themselves (such as to make it a meaningful modulator of influencers' responsibility) it would be necessary for disingenuousness or manipulative intent or some other intentional stance on the part of influencers to *directly* impact the truth and relevance of what is shown in their influencings. This seems impossible.

An advertiser, say, may disingenuously claim 'our shoes make you run faster' while not believing said claim. If their shoes – as it happens – *do* make one run faster, then the problem is with the advertiser, not the advert. The advertiser acts disingenuously but, by chance, harms nobody. This doesn't mean he was making adverts ethically; in my view he wasn't and on account of this we might plausibly call him 'manipulative[395]' and blame him for disregarding the norms of ethical influencing despite getting morally lucky. This is different, however, from regarding the influence itself as bad (or at least worse) and holding the influencer responsible for the influence's badness, such as would be needed for influencers to be meaningfully responsible for the effects of their influencing in cases where they influence only by showing reasons.

The escape from this argument, for anyone proposing some sort of intention as a modifier of influences' legitimacy, is to hold that some intention is somehow in-itself able to account for a change in the ethical qualities of an influence carried out with said intention. No such view can

---

[393] R.Noggle (2020), 'Manipulation: A Unified Account', *American Philosophical Quarterly*, vol.57, no.3, pp.251-252; pp.243-244

[394] Generally, one might still be responsible for one's character as an influencer even though, as it happens, one is not responsible for the effects of one's influencing (on dispositions or on things causally downstream from effected dispositions) because this influencing is accomplished through showing reasons. One might hence still be blamed (or indeed praised) for this character even where one cannot be blamed for the effects of one's influencing.

[395] C.Mills (1995), 'Politics and Manipulation', *Social Theory and Practice*, vol.21, no.1, pp.97-112; p.100, R.Noggle (1996), 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly*, vol.33, no.1, pp.43-55; pp.49-50

succeed, however, because no such views are able to plausibly account for the occurrence of evaluatively distinct accidental influences.

Dave is hit by a falling rock in a natural landslide. The rock, by pure chance, hits Dave precisely such that it destroys Dave's existing disposition to give to charity by bruising exceptionally specific parts of Dave's brain, forcing the re-arrangement of a very particular set of neural pathways concerned with charitability. Enid is caught in a terrible car accident. She loses a leg when she is hit by a drunk driver. Enid is moved in such a way by the whole experience that she quits her job and begins advocating against drunk driving.

Both Dave and Enid have, I suggest, been subjected to influences which have had effects on their moral dispositions; Dave has been made less charitable, Enid an advocate against drunk driving. These influences had no intent behind them; they were (let's say) pure accidents, happenstance. Nonetheless, I suggest, it makes sense to think about whether these influences were or were not acceptable – are they the sorts of changes in people we should put up with in life, or take special measures to prevent and resist? Is it an evaluatively inert fact that Dave's accident made him less charitable, or is it part of the tragedy of the accident that it had the effect on his moral dispositions that it did? Isn't it okay – or at least better – that Enid was moved by her experience to change her moral dispositions, in a way that it wasn't for Dave? Isn't there something about getting into road safety following an accident that seems consistent with proper human agency in a way that changing one's charitability following an accidental blow to the head isn't?

When one evaluates influences according to the intentions behind them one blinds oneself to such nuances. One must say that we should view the changes visited on Dave and Enid in the same way, for they both result from properly unintentional events. One must thus either deny the possibility of 'manipulation by circumstance', say by requiring a specific intention for unacceptable influence (like Yaffe seems to[396]), or else hold that all influences originating in accidents should be rejected (if one requires a specific intention – like an intention to inform – for acceptable influencing). Availing of the former approach one sacrifices an intuitive part of the tragedy of Dave's story; that he was changed in an unacceptable way. Availing of the second approach the intention-theorist loses any ability to distinguish between the acceptability of Enid's change and the acceptability of Dave's; one (on such an account) cannot understand Enid's change as a rationalizable response to events in a way that Dave's isn't. Either way, evaluating influences in terms of the intentions behind them discards intuitive detail.

Yaffe, using a language of freedom, resists this point by suggesting that those who are accidentally influenced are afterwards left to change behaviour freely, whereas those subject to influence with intent will be subjected to ongoing campaigns of control which will 'track' their changing dispositions in order to ensure particular outcomes[397]. As Yaffe admits, though, it's possible for intentional influences to fail to control agents over the long term and accidental influences to succeed in doing so[398]. In admitting this, however, Yaffe admits the incidental rather than essential part intention plays in his analysis. Yaffe's contention is not really that a specific intention is necessary for illegitimate influence but rather that a specific intention – the intention to reduce an agent's available future options – makes one more likely to exert illegitimate influence. This seems true;

---

[396] G.Yaffe (2003), 'Indoctrination, Coercion and Freedom of the Will', *Philosophy and Phenomenological Research*, vol.67, no. 2, pp.335-356; p.344
[397] *ibid*. p.344
[398] *ibid*. p.346

wanting an agent's behaviour to change in a very specific way might lead one to abandon scruples about how one goes about influencing them.

Such a lack of scruples, however, isn't a necessary feature of the influencer with clear ends in mind and even an unscrupulous influencer might employ unproblematic means of influencing when they're effective. A disingenuous or otherwise ill-intentioned influencer may be relatively more likely to exert illegitimate influence, caring relatively less about how they go about influencing, but it isn't such a character's intentions themselves which make their influencing problematic. Such intentions merely make them problematic, precisely because they seem more likely to influence in unacceptable ways whether-or-not they in fact do so.

This part intentions can play in evaluating influences leaves no room to connect the intentions of influencers to the effects of influences such that they bear responsibility for them, at least insofar as influencing is done by showing reasons. A rabble-rouser might intend to see his enemy destroyed and spread stories about them so that others hate and destroy the enemy. If all these stories are true and relevant to whether the enemy should be hated and destroyed (suppose the enemy is such that there are some strong – if not necessarily sufficient – pro-tanto reasons to hate and destroy them), though, I suggest that we shouldn't hold the rabble-rouser responsible for the enemy's destruction – we should hold those who actually destroyed the enemy responsible, perhaps the enemy themselves to some extent (if, say, they created reasons for people to hate and destroy them). We might perhaps hold the rabble-rouser responsible for his intentions and we might blame him for them – if we think, say, that to intend to create hatred is wrong, or if the rabble-rouser rabble-roused in bad faith – but these are different things.

Thus, I argue, the person who effects changes in behaviour by revealing reasons isn't responsible for changes in agents' dispositions in response to the reasons he reveals.

There are multiple ways to make sense of this claim in using broader understandings of the nature of responsibility. We might propose, with some[399], that human agency breaks causal chains that would otherwise pass along responsibility. This proposal doesn't sit well with my account, inasmuch as my account is dependent on tracing causal chains through human agency (for 'influences', as I discuss them, are causes of change in the behaviour *of* human agents). We might say, rather, that human agency breaks chains of responsibility, such that when a causal chain passes through steps which are proper parts of human agency (such as psychological processes triggered by exposure to reasons) agents backwards in the causal sequence don't acquire responsibility for events further forwards. When you show another a reason to act differently – for whatever reasons of your own – how their dispositions adjust to it isn't your responsibility.

To show someone a reason then, even compellingly and with the intent to trigger a specific change in their behaviour, isn't to acquire any responsibility for their change in behaviour. This responsibility rests with the person themselves and perhaps in some sense the reason shown (at least, it is to this reason that the agent should react and not you as the person who showed them it, if it indeed be a reason). As I'll argue in the following section, this isn't the case when one exerts influence using means other than showing reasons. It also shouldn't be interpreted as claiming that reason-showers can't do wrong by showing reasons. They just shouldn't be considered responsible for whatever influence over dispositions the reasons they show wield.

(5.10) Mediating Reasons, Using Non-reasons

[399] R.W.Sellars (1957), 'Guided Causality, Using, and "Free Will"', *The Journal of Philosophy*, vol.54, no.16, pp.485-493; p.490

If it's true that those that show reasons and thereby exert influence aren't responsible for the influence 'they' exert in such a way, it's hard for the proponent of interpersonal freedom to directly object to influences my account finds unproblematic. Before this point may be argued, though, I must first resolve a semantic problem with my position here.

I've claimed, in essence, that influencers who influence only by showing reasons don't influence. Only the reasons they show influence in such cases. In what sense, then, may such characters usefully or properly be called 'influencers'? If they aren't influencing what are they doing, and what is the sense in addressing this activity – whatever it is – in the context of a thesis trying to offer standards for evaluating methods for influencing moral dispositions?

We have a conventional understanding able to resolve this apparent problem, though its application here must be explained. The journalist, in particular, is called a member of the 'media', a 'mediator' of relevant truths rather than a user of them. This is an understanding consistent with my position. On this understanding, the journalist should be construed not as a person who exerts power over others by showing them reasons, but as a *medium* for whatever impersonal influential power is more-or-less (depending how you understand the nature of reasons) inherent in such reasons (just as a tape recording may act as a medium for the impersonal influential power of a particular sonic event, such as some beautiful birdsong that might inspire one to help protect birds).

This understanding applies so long as the journalist is doing conventionally good journalism – so long as they publicise relevant truths, without the addition of falsity or extraneous sources of influence (such as, for instance, irrelevantly emotive music or rhetorical devices to amplify the aversions or attractions evoked by the story being told)[400]. When such things are added the journalist becomes other than a mediator; they become at least partly responsible for the influence of their output insofar as at least some of said influence is produced by the journalist's industry rather than merely exposed by it.

Even when such things are not added, however, and reasons assume responsibility for their influencing whoever shows them, it would be a mistake to disregard the capacity of reasons' mediators to direct their influence in various ways. A good-but-activist journalist, say, or honest propagandist may still use reason-showing as a method for achieving the changes in moral dispositions that are their objectives, even if in so doing they only make themselves and their work into mediums for the influence of the reasons that they show.

My point here is that showing reasons to agents to change their moral dispositions has a special character as a method of influencing, but it may nonetheless be called a method of influencing. Per this character the influence in such cases shouldn't be seen as produced or in the power of an influencer but rather as something found and mediated by an influencer. This mediation may helpfully be called a method of influencing, but it is a method where it isn't really the 'influencer' doing the influencing; the influencing is being done by relevant facts beyond the influencer's designs. These facts would be no less influential – though they may be less well publicised or appreciated – were the influencer uninvolved. The influencer (who is properly a 'would-be' influencer in such cases) mediates the influence of these facts but doesn't thereby participate in or assume responsibility for their influencing. The would-be influencer in such cases doesn't exert influence, though they aid things which do.

[400] S.Tait (2011), 'Bearing Witness, Journalism and Moral Responsibility', *Media, Culture & Society*, vol33. no.8, pp.1220-1235; p.1232

This sort of legitimate influencing by mediation must be distinguished from the mediation of illegitimate influences. My argument can defend the former from attributions of illegitimate interpersonal power, it can't similarly defend the latter.

The mediator of reasons is a mere mediator because the things she mediates are such that none should be held responsible for their influence, provided at least that their influence is working properly (i.e. provided that the influenced agent is not defective as a converter of reasons into dispositions like the 'Eiffel Tower eater' of §4.5), for under such circumstances they assume this responsibility (perhaps also pass it on to those who are influenced by them). Reasons, I've argued, assume responsibility for any influence they exert upon dispositions (that isn't passed to influenced agents themselves) because it is to only them we should respond in changing our dispositions[401]. Things other than reasons – irrelevancies – are not such that we should respond to only them in changing our dispositions. Thus, such irrelevancies can support no similar argument liberating influencers who employ them from responsibility for what seems to be their influencing.

Thus, we've no reason to regard those who mediate or produce influences which don't work by showing relevant reasons as free of responsibility for changes in dispositions which result from their outputs' influencing. If such characters – dishonest propagandists, for instance – by their mediation or production of non-reasons manage to acquire any influence this influence can amount to problematic interpersonal power in a way that the influence of relevant reasons cannot.

Think back to my example of the journalist who reveals that there's pollution in the water supply. Suppose a variant case where the journalist in question is more like a dishonest propagandist (and a conventionally awful journalist). Said journalist achieves the same changes in dispositions – scares people away from the public water – but does so by telling a compelling but thoroughly vacuous story, containing no genuine reasons to avoid the public water but rather a mix of fabricated images, effective rhetoric and emotive music.

What would it be to look through said journalist's method of influencing and hold the relevant truths within it responsible for their effects upon behaviour? There are, in a case like this, no relevant truths available to hold to account; no pollutants to remove from the water and no polluters to blame (and if there are, by fluke, any such things they aren't represented in the journalist's story). The emotive music used in the journalist's reporting, say, is a real thing but not a real thing which one ought to hold responsible for its effects on dispositions, precisely because it isn't a relevant thing with the right character needed to seize such responsibility. As there is nothing in the influences deployed by such a 'journalist' that we ought to hold responsible for what seems to be *their* influencing we should regard this influencing as something done by them themselves, for they were responsible for the causal chain behind it. We should thus regard the journalist in this case as responsible for the changes in dispositions that resulted from their story and thus plausibly as a holder of possibly unjust interpersonal power with respect to those influenced.

This analysis applies to some extent to any person involved in the exertion of influence through anything other than the presentation or representation of relevant facts to proper reason responders – from the unscrupulous propagandist to the workman who pastes such propagandists' messages to walls to the lobotomist who cuts at brains to create docility[402]. All such agents are

[401] Though there are cases where we should – in a sui generis sense – respond to other things in order to secure lesser-evil results. This is consistent insofar as it may be the case that one should X and that one sui generis should not X for one must not X in order to Y and Y-ing is better than X-ing.

[402] A.Tone and M.Koziol (2018), '(F)ailing women in psychiatry: lessons in a painful past', *Canadian Medical Association Journal*, vol.190. is.20, pp.624-625; pp.624-625

properly responsible, more-or-less, alone or alongside others (and for better or for worse), for the changes in dispositions caused by their endeavours. This contrasts with those who achieve such changes by showing reasons, who are not responsible for the dispositional changes caused by their endeavours. Those who achieve dispositional changes through a mix of reason showing and other methods, meanwhile, risk responsibility for the effects of their influencing to the extent that these effects depend upon their use of such 'other methods'.

(5.11) Responsibility and Interpersonal Freedom

I've argued that it's wrong to hold those that use the revelation of relevant truths – reasons – as a method of influencing responsible for whatever influence they seem to wield. Rather in an important sense the relevant truth itself assumes responsibility for its influence when this influence is secured by its presentation or representation to agents. This is the case because where the relevant truth moves us to action this action should respond only to this truth rather than the means and intent by which it's revealed to us. This is not the case, however, with respect to influencing using methods other than the revelation of relevant truth.

This point, I suggest, should make a difference in how the advocate of interpersonal freedom thinks about influencing by showing the relevant truth. It implies, I argue, that such an advocate ought to endorse (at least in practice) my rationalizability requirement.

Recall the honest propagandist as I introduced her, prima facie problematic to the proponent of interpersonal freedom yet unproblematic given the rationalizability requirement. What made this character problematic for the proponent of interpersonal freedom was her apparent power over the dispositions of those subjected to her propaganda. She seemed, by her effective (though honest) propagandising, to be able to take charge of the actions of those subjected to her propaganda by showing these subjects vivid reasons and thus exerting influence over their behaviour. Such a relation seemingly threatens interpersonal freedom – it makes the actions of one subject to the interference of another (which would, say, trouble Berlin[403]), even a 'domination' plausibly undisciplined by the interests of the interfered-with agent (which would, say, trouble Pettit[404]).

If I'm right in my analysis of what and whom we should hold responsible for the influence of shown reasons it would be wrong to hold the honest propagandist responsible for whatever influence her reason-showings have on moral dispositions. Agents' proper responses to reasons vary only with their contents and not with who shows them reasons or any features of reason showers; the reason-shower like the honest propagandist thus cannot be meaningfully held responsible for 'their' influencing of dispositions. Reasons themselves in some sense or perhaps influenced agents (insofar as they can control their responses to reasons[405]) must instead be held responsible. If any power is wielded over influenced agents in cases like that of the honest propagandist, thus, where influencing proceeds only through agents' reactions to shown reasons, this power must thus be wielded by either relevant facts or influenced agents themselves.

Either way, interpersonal freedom should be preserved despite the influence of either thing upon one's dispositions. A relevant fact is not a person. It may involve people, but it itself cannot be a person. It's never other than a state of affairs which recommends, constitutes reason for, certain

---

[403] I.Berlin (2002 [1957]), 'Two Concepts of Liberty' in H.Hardy [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217; pp.169-170

[404] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; pp.63-64

[405] *see* C.Korsgaard (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press; pp.164-166

dispositional change(s). As a non-person a relevant fact, no matter how compelling, can never pose a threat to any sort of interpersonal freedom. It can't be a somebody else exerting influence over and dominating your actions because it can't be a somebody else[406]. Equally oneself cannot be a somebody else dominating one's own actions for oneself is never a somebody else.

Interpreting this thought given Berlin's interpersonal account of freedom, the compelling relevant truth may be conceptualised as a non-person determinant of one's range of action analogous to a natural material condition (for example, a natural resource scarcity or glut). Just as such limitations cannot remove one's interpersonal freedom for Berlin[407] the compunction of relevant truths cannot strip one of interpersonal freedom. Similarly, for Pettit, absent external agency one cannot have domination of the agent and a threat to the freedom he valorises[408].

Common to these understandings is the idea that the influence of reasons isn't an external agent forcing its will upon one. It's more like a constitutive part of one's own agency in a particular context, at least when it may be held responsible (in the special sense that such a thing may be held responsible) for whatever effects it has on dispositions. Pettit suggests rational control seems to be necessary (if not sufficient) for an agent to be free[409]; it could hardly be a thing inconsistent with freedom. Berlin, similarly, suggests one's individual freedom consists in, at least partly, being 'moved by reasons … which are my own[410]' (by which he seems to mean reasons which one finds compelling[411], an understanding consistent with the rationalizability requirement).

I suggest the following: interpersonal freedoms demand that one's choices not be interfered with (perhaps excepting certain circumstances[412]) but (at least superficially) leave open the question of what may properly constitute one's choices. My position is that under all circumstances the influence of relevant facts upon behaviour (where there is such influence involved in determining one's choices) is amongst the things that properly constitute one's choices. Responsiveness to such influences in one's dispositional development is part of proper agency. This principle holds no matter the genealogy of such influences, including who shows you them and why.

Consistent with the historic intent of conceptualising freedom in interpersonal terms[413], this understanding doesn't justify one being 'forced to be free' by being coerced into acting as reasons say one should. What matters on my account is not that one is influenced '*for a good reason*' (as Aylsworth says[414]) but *by a good reason*. The compunction my account valorises is only whatever compunction may be supplied by the presentation or representation of reasons. Forcing someone to

---

[406] This doesn't mean that facts can't limit one's freedom in various ways – armlessness may compromise one's freedom to arm-wrestle – but this I here speak to *interpersonal* senses of freedom in particular; ones concerned only with power relations between persons.

[407] M.Dimova-Cookson (2013), 'Defending Isaiah Berlin's Distinctions between Positive and Negative Freedoms' in B.Baum and R.Nichols [eds.] (2013), *Isaiah Berlin and the Politics of Freedom: 'Two Concepts of Liberty' 50 Years Later*, New York, Routledge, pp.73-126; p.107

[408] P.Pettit (2006), 'Freedom in the Market', *Politics, Philosophy and Economics*, vol.5, no.2, pp.131-149; p.132

[409] P.Pettit (2001), *A Theory of* Freedom, Cambridge, Blackwell; p.48

[410] I.Berlin (2002 [1957]), 'Two Concepts of Liberty' in H.Hardy [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217; p.178

[411] *ibid*. p.178

[412] P.Pettit (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press; pp.63-64

[413] *see* S.Colingnon (2018), 'Negative and Positive Liberty and the Freedom to Choose in Isaiah Berlin and Jean-Jacques Rousseau', *The Journal of Philosophical Economics*, vol.12, no.1, pp.36-64; p.39

[414] T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; p.695

follow the reasons you believe they have for action will involve the application of other forms of influence, such as the threat of harm, and will violate the rationalizability requirement.

(5.12) Interpersonal Freedom and Relevant Truth

The influence of reasons, then, cannot threaten interpersonal freedom, for reasons can't be people trying to control you. They may sometimes be mediated by people trying to control you but these situations never justify us regarding the influence of reasons thus conveyed as somehow threatening to interpersonal freedom. Interpersonal freedom consists in making your own choices and your responses to reasons – however you come to perceive them – are always a proper part of the *you* making your choices.

If this is the case, what does this imply about accounts of interpersonal freedom themselves? How do they relate to the rationalizability requirement?

In §5.7 I suggested that the proponent of interpersonal freedom has prima facie grounds to dismiss my position. They might argue the rationalizability requirement valorises interference in the behaviour of agents through showing reasons. If my response to these worries holds, however, there can never be such 'interference'; being influenced by the relevant truth cannot constitute an interference in one's behaviour no matter the origins of one's influencing. This implies that the proponent of interpersonal freedom's prima facie grounds for dismissing my position don't hold up to scrutiny.

Applying this insight to practical cases yields a particular view of how we should react to certain influences. For example, consider popular concern about public information campaigns, which worry they distort peoples' ability to make free choices (see §5.7). If I'm right, then a significant test of whether such campaigns actually can distort free choices is the rationalizability requirement: do these campaigns present or represent relevant reasons? If they do – and do nothing else that has an influence on dispositions – then they cannot directly limit the ability to make interpersonally free choices.

Granted, in certain circumstances such campaigns (and analogous influences) might indirectly damage interpersonal freedom. Suppose, say, the campaign shows reasons to drive less and hence influences one towards driving less without compromising one's interpersonal freedom. Suppose also that people who drive less are less interpersonally free – suppose they systematically spend more time stuck in the company of tyrannical relatives. If this were the case, then we might say the information campaign *indirectly* compromises interpersonal freedom by – in a way compatible with interpersonal freedom – influencing one to reduce one's interpersonal freedom. The problem in such a case is not how one is influenced but what one is influenced to do and the rationalizability requirement will never protect one's interpersonal freedom in such cases (provided, of course, that there can be pro tanto reasons to take actions which restrict one's interpersonal freedom).

In such and all other cases, however, the rationalizability requirement remains useful to the proponent of interpersonal freedom. Irrespective of the effects that the results of an influencing may have on one's freedom one should still ask whether being influenced itself might compromise one's freedom. For the proponent of interpersonal freedom the rationalizability requirement, if I'm right, can help answer such questions. By separating cases where responsibility for one's influencing may be imputed to other agents from those where it may not be, it can help us detect which influencings can threaten interpersonal freedom and which cannot. Thus if you value freedom in interpersonal terms, like Berlin or Pettit, then I don't think you can escape from the need to embed the rationalizability requirement, somehow, in your evaluations of influences. You must, for

example, classify compunctions triggered by reasons – no matter how they reach one – as necessarily unproblematic constitutive parts of one's agency, rather than as potential means of interference in one's choices. This makes a difference in one's position.

To return to some examples highlighted in §5.7 this means that the proponent of interpersonal freedom cannot complain (in terms of threats to people's ability to make free choices, at least) about – say – a message on the side of a cigarette packet reading 'smoking kills' provided smoking does in fact kill and this representation of the lethality of smoking is relevant to those who'll encounter it. They also cannot complain (again, citing threats to peoples' ability to choose freely) about the leaking of information which influences behaviour provided the information is true and, again, plausibly relevant to those it's leaked to.

Importantly, I think I've shown that such points hold irrespective of who supplies the reasons in question and their intentions. Whatever power such agents think they secure is illusory; when we change behaviour in response to relevant facts we do so subject to the compunction of reasons, not people. This isn't to say that who tells us things doesn't matter but rather that it matters only when we're assessing the truth and relevance of what people say. Reasons' compunctions are never somebody else's tools for tyrannising us, mediators notwithstanding. The proponent of interpersonal freedom, like the proponent of autonomist freedom, has reason to endorse the rationalizability requirement.

(5.13) Conclusion

Extant debate about the legitimacy of influences has stressed the importance of the preservation of freedom through processes of influencing. In this chapter I've shown that on two of the most influential understandings of freedom – freedom as autonomy and freedom as interpersonal non-domination – this preservation may be achieved by the application of the rationalizability requirement. Either autonomous agency requires de dicto or de re motivation by reasons, if the former then this agency loses attractiveness while if the latter then this agency may be protected by the rationalizability requirement. Interpersonal domination through influencing, meanwhile, requires that one be influenced by another agent. Reasons aren't agents; insofar as they assume responsibility (perhaps along with oneself) for one's influencing in accord with the rationalizability requirement (as they do) one cannot directly lose any interpersonal freedom when one is influenced in accord with the requirement. Given all this, proponents of autonomist and interpersonal freedom have reason to embrace the rationalizability requirement. If one regards either of these freedoms as worth preserving through processes of influencing one ought to take the rationalizability requirement seriously as an ethical constraint on influencing (at least moral dispositions).

Chapter 6: Conclusions

(6.1) Introduction

I've argued for a particular perspective on the ethics of influencing human moral dispositions, one which proposes that, ceteris paribus, we should be influenced by things which present or represent reasons and that deviations from this ideal stand in need of justification. Acting upon dispositions through the representation or presentation of relevant reasons is always a better-making feature of an influence. I've shown how this perspective may be used in practical judgement, that it coheres with the most plausible interpretation of the autonomy tradition and that it can't valorise problematic power relations. In this concluding chapter, building upon these reflections, I'll try to say what my perspective – properly applied – may mean for some existing practices and debates. I'll speak to the implications of my perspective for the evaluation of advertising methods, methods of moral bioenhancement and 'nudges', as well as the understanding of certain evaluative vocabulary. In doing so I intend to say something about what we, confronted with ever-growing powers to manipulate human moral behaviour, ought to do.

(6.2) An Ethics of Moral-dispositional Change

I'll begin by recapping my position.

This thesis is meant to capture the content distinctive to a certain part of ethics; the part encountered only in asking questions like 'should I tolerate this influence on my behaviour?' or 'how should I make this person act the way I want him to?' where the dispositional changes involved have a 'moral' character.

This domain limits the scope of the principles defended in this thesis. These principles govern only the evaluation of changes in distinctively moral dispositions, those which determine behaviours with 'moral' characters as defined in §2.2. They say nothing direct about the influencing of other dispositions, direct control over behaviour itself or the influencing of other parts of psychology (such as, for instance, non-moral beliefs or habits). They also cover only that part of the totality of ethics that might apply to our judgements about the acceptability of such changes which is distinctive and general to thinking about such changes.

Within this scope, however, the principles defended in this thesis have much to say. They say that there exists, setting other normative considerations temporarily aside, a line between methods of changing moral dispositions which are in virtue of their characters 'legitimate' and methods of changing moral dispositions which are in virtue of their characters 'illegitimate'. Legitimate methods of changing moral dispositions, while they needn't make us better people, are such that their action upon our dispositions is better in a significant way, understandable at least in terms of the preservation of valuable reason-respondingness in our behaviour and perhaps also certain freedoms. Illegitimate methods, by contrast, while they needn't make us worse people, are systematically worse sorts of influence in the same way; insofar as they're effective, they alienate dispositions from those realities that ought to shape them and perhaps risk certain freedoms.

Things are this way because, I argue, there's value in at least moral dispositions being formed in response to experience (either through successful representation or presentation) of relevant parts of reality, reasons, which in virtue of their characters are such that they more acceptably compel their experiencers to undergo certain moral-dispositional changes. This is, I suggest, the best

explanation of intuitions about cases like AB. Given this explanation for an influence N to induce dispositional change C legitimately (in a properly functioning reason-responder), with respect to the ethics distinctive to influencing moral dispositions, two things must be the case. First, N must represent or present some justifying reason to make change C. Second, this representational or presentational content must be causally responsible for the influence exerted by N upon C. There's a rationalizability requirement on changing moral dispositions.

Any influence upon moral dispositions which achieves this rationalizability deserves a certain respect and special treatment as a fitting inspiration for moral-dispositional change; the sort of thing that ought to change one. As such, such an influence ought not be opposed on account of the sort of influence it is, and there exists a burden of ethical proof upon anyone who argues such an influence ought to be resisted for other reasons. This burden may be carried, for example when grave consequences are somehow in play (for the demands of the rationalizability requirement are only pro tanto), but it should not be ignored by someone who has designs upon the changing of moral dispositions.

Any influence upon moral dispositions which fails to attain this sort of legitimacy, by contrast, demands resistance and avoidance as a threat which usurps the proper role of experiencing relevant reality in the determination of how we tend to act in moral contexts. Again, such demands may be overruled by others, but this still means that the proponent of any such 'illegitimate' method of influencing must shift a burden of ethical proof if he's to make his point convincingly.

In any particular case, the strength of all positive and negative demands of the rationalizability requirement will be a function of how effective the methods of influencing involved are, how important the dispositions being changed are, and how many will be influenced by whatever influences are at issue.

The respectability of these demands varies only with the causal role that showings of reasons play or fail to play in the influencing of moral dispositions. Contra some versions of the autonomy tradition[415], legitimate influencing needn't require cognitive mediation between reasons and dispositions (though it does require experiential mediation of influences' relevant content). Whomever or whatever orchestrates influencing, whether it be the influenced party, their enemy, or a random accident, also counts for nothing.

These demands and the standards behind them, I think, capture what's distinctive to the ethical evaluation of influences on moral dispositions. Fortified by this analysis, it should now be possible to answer the questions which motivated this thesis. It should now be possible to say something more about what we, aware of power over our own and others' lived morals, ought to do.

(6.3) Defining Brainwashing

Explicit, richly articulated concerns about unethical means of influence are a relative novelty within secular conversation (one may find precursor debates about the Christian concept of grace[416]). They go back more-or-less seventy years[417], not-coincidentally to a period of innovation in the tools of

[415] *see* A.M.Baxley (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press; pp.30-34
[416] T.W.Cyr and M.T.Flumer (2018), 'Free Will, Grace and Anti-Pelagianism', *International Journal for the Philosophy of Religion*, vol.83, no.2, pp.183-199; pp.183-184
[417] M.Introvigne (2014), 'Advocacy, Brainwashing Theories and New Religious Movements', *Religion*, vol.44, no.2, pp.303-319; p.304

propaganda and advertising facilitated by advancements in the understanding of the mind[418]. Such concerns often make use of the mid-20th century neologism[419] 'brainwashing' and related terms such as 'manipulation'.

What, though, is this thing called 'brainwashing', or at least, what more can be said about what it is given the rationalizability requirement?

Prima facie what's clear about the term 'brainwashing' is that it's 'thick'; a term like 'murder' loaded with evaluative meaning. To murder someone is to do more than kill them; it's to kill them in a context which makes said killing wrong given standards which specifically govern the ethicality of killing people. A murder doesn't become a murder just because it's a sui generis wrong killing, or a killing that's wrong because the killed person was going to save ten lives tomorrow. It can only become a murder if it is made such by standards which *specifically* determine the wrongness of killings, as opposed to actions more generally. Similarly, to brainwash someone is to do more than influence them; it's to influence them in a way which is wrong in virtue of standards that specifically govern the ethicality of influencing people. Brainwashing isn't just wrong influencing; it's influencing in a way that violates standards of ethical influencing.

The rationalizability requirement I've articulated comprises some such standards (perhaps not all of them) and this being the case it can furnish part of an explanation of what it is to brainwash or be brainwashed and indeed of why such things are troubling.

'Brainwashing', generally, equivocates between referring to violations of norms governing the changing of ideals and violations of norms governing the changing of dispositions. Sometimes we're 'brainwashed' into thinking differently (into believing we should behave differently, irrespective of how we indeed behave) and sometimes we're brainwashed into behaving differently; sometimes both. The side of this equivocation concerning ideals is the territory of the epistemologist, and beyond my account.

On the dispositional end of this equivocation, the term may apply to violations of norms specifically governing disposition-changing in general or only change in moral dispositions. I'm not sure whether there are any norms specifically governing disposition-changing in general (I certainly haven't tried to prove that there are). There are, however, I've argued, norms specifically governing the changing of moral dispositions; these are those norms that constitute the rationalizability requirement.

I suggest these norms can define an important subset of the ways in which 'brainwashing' can happen. Wherever moral dispositions are changed by things other than presentations or representations of reasons relevant to the changing of said dispositions brainwashing – to some extent – occurs. The wrongness of this brainwashing, in turn, is to be explained by its diversion of the 'brainwashed' agent from a better sort of agenthood which requires moral dispositions which result from experiencing relevant parts of reality.

This analysis extends, somewhat, to cognates of brainwashing. Manipulation, notably, insofar as it may be understood as meaning the same thing as brainwashing (which it sometimes may, though some argue manipulation always requires certain sorts of intent[420]), may be similarly interpreted

---

[418] M.E.Drumwright and P.E.Murphy (2009), 'The Current State of Advertising Ethics: Industry and Academic Perspectives', *Journal of Advertising*, vol.38, no.1, pp.83-108; p.84

[419] M.Holmes (2016), 'The 'Brainwashing' Dilemma', *History Workshop Journal*, vol.81, no.1, pp.285-293; p.286

[420] *see for examples* C.Mills (1995), 'Politics and Manipulation', *Social Theory and Practice*, vol.21, no.1, pp.97-112; p.100, R.Noggle (1996), 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly*, vol.33, no.1, pp.43-55; pp.49-50

using my account. At least, the pejorative aspect of the term (to manipulate someone is conventionally to do them wrong in some way) may be helpfully analysed using the rationalizability requirement. What's bad about manipulating moral dispositions may be understood, given the requirement, as (at least in some cases) being such manipulation's diversion of the manipulated party from the kind of behaviour formation they ought to undergo.

The analysis here also helps us address at least some puzzles associated with brainwashing.

Why, for instance, are our intuitions about which act-types constitute brainwashing strongest when we think of mature moral agents as opposed to moral agents which are underdeveloped (such as young children) or defective (such as those with certain psychopathologies)?

For many given brainwashing-apt methods of influencing M, we tend to be more willing to visit M upon underdeveloped or defective moral agents than mature and functioning ones. For instance, repeating association-laden messages ('your favourite cartoon character does things this way') in order to mould behaviour is a fairly standard parenting technique but used against mature individuals it's regarded as a propagandist's 'dirty trick[421]', in the vicinity of brainwashing[422] (though one might prefer to reserve this term itself for extreme applications of such techniques). Use of pharmaceuticals with the potential to influence moral dispositions on those with psychiatric illnesses is standard medical practice[423] but is controversial when applied to those without such illnesses[424]. This class of intuitive evaluative differences runs counter to others which cover similar territory; usually wronging a young child or someone with psychological problems is regarded as worse than wronging a healthy adult in the same way. With brainwashing this pattern can become inverted (this is not to say that 'leading someone astray' seems less problematic with regard to such agents – quite the opposite – but that leading the person using bad methods – brainwashing – generally seems less problematic).

My analysis explains this inversion by connecting brainwashing, or at least the subset of brainwashings it may be used to define, to an idealised moral agenthood. It is on account of the value of being reason-responders, contingent on our capacity to be such, that we ought not to be 'brainwashed' (at least insofar as the reality requirement may ground attributions of brainwashing). Those who – at least for now – lack this capacity cannot be such that they ought not to be brainwashed in the same way. Thus it is that such agents – young children and people with certain psychologies (think, say, of a psychopath uncompelled by human suffering[425] and thus plausibly defective as a reason-responder) – are such that they may be acceptably subjected to moral-dispositional influences which would cross the line into brainwashing if used on other agents.

Thus, I argue that the definition of illegitimate influences on moral dispositions encoded into my account contributes usefully to (though doesn't exhaust) identification and understanding of the

---

[421] P.Biegler and P.Vargas (2016), 'Feeling is Believing: Evaluative Conditioning and the Ethics of Pharmaceutical Advertising', *Journal of Bioethical Inquiry*, vol.13, no.2, pp.271-279; pp.275-276

[422] J.T.Richardson and M.Introvigne (2001), '"Brainwashing" Theories in European Administrative and Parliamentary Reports on "Cults" and "Sects"', *Journal for the Scientific Study of Religion*, vol.40, no.2, pp.143-168; pp.154-155

[423] N.Levy, T.Douglas, G.Kahane, S.Terbeck, P.J.Cowen, M.Hewstone and J.Savulescu (2014), 'Are You Morally Modified? The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-171; pp.118-123

[424] J.Specker, M.H.N.Schermer and P.B.Reiner (2017), 'Public Attitudes Towards Moral Enhancement. Evidence that Means Matter Morally', *Neuroethics*, vol.10, no.3, pp.405-417; pp.405-416

[425] E.Ramirez (2016), 'Neurosurgery for Psychopaths? The Problems of Empathy and Neurodiversity', *AJOB Neuroscience*, vol.7, no.3, pp.166-168; pp166-167

troubling activity we call 'brainwashing'. It tells us why we call certain things – moral-dispositional influences effected without the showing of relevant reasons – 'brainwashings' and it tells us something about why these things seem troubling, locating their problematic character in their alienation of agents from the reasons that ought to inspire their moral conduct.

(6.4) The Rationalizability Requirement and Advertising Ethics

In this thesis I've repeatedly considered the acceptability of various advertising methods. I've asked, in one form or another, whether there are methods of advertising which are unethical in virtue of their characters as influences and what explains this status.

There is an extant literature challenging the acceptability of advertising methods, either specifically as influences as with my account or indeed in sui generis terms[426]. While this is not the right place to discuss in detail particular contributions to this literature, I do suggest that the insights offered by my account have implications for advertising ethics (at least wherever adverts concern behaviour which has moral significance).

Consider that my account maintains a sort of psychological neutrality. It demands inspiration by perceived relevant reasons for legitimate moral-dispositional influencing, but it doesn't valorise any very particular psychological connection between these reasons and dispositional changes. It (following, say, Arpaly[427]) doesn't require, notably, that reasons shown by legitimate influences act upon our beliefs and only through them effect dispositions. To demand this would be to demand, as I've argued (see §3.8), that behavioural change be triggered by those representations we call beliefs rather than the parts of the world that ought to influence our behaviour themselves. This gets things wrong. Relevant perceived *content*, and its causal responsibility for dispositional change, is sufficient to (in a distinctive pro-tanto way) legitimate the influential effect of an influence. My account doesn't care whether the causal mediators between this content and dispositions consist of beliefs, emotional states or neurotransmitters (provided that these causal mediators don't somehow produce a separate influence in-themselves, say by triggering some sort of self-influencing behaviour[428]). Provided mediators properly connect perceived content to dispositions in whatever case or set of cases is at hand by being such as to reliably generate changes in dispositions appropriate to said content valuable rationalizable influencing can take place.

This neutrality, in turn, allows my account to assess the legitimacy of methods of influencing used in adverts, whatever psychological levers these methods pull. This contrasts with conventional autonomist analyses which may dismiss many advertising techniques out-of-hand on the basis that they work non-cognitively[429]. What matters – contra such analyses – is that behaviour is inspired by appropriate reasons, not which representations convey the inspiration or whether these representations are such that they may be called beliefs. Incorporating this insight allows for a stronger and more nuanced evaluative project.

What impact does this 'nuance' have?

---

[426] M.E.Drumwright and P.E.Murphy (2009), 'The Current State of Advertising Ethics: Industry and Academic Perspectives', *Journal of Advertising*, vol.38, no.1, pp.83-108; pp.84-85

[427] N.Arpaly (2002), 'Moral Worth', *The Journal of Philosophy*, vol.99, no.5, pp223-245; pp.238-239

[428] In such a case it would be necessary to disambiguate the at least two influences involved and assess them separately using the rationalizability requirement.

[429] T.Aylsworth (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699; pp.692-693, *also see for example* H.C.Brown (1929), 'Advertising and Propaganda: A Study in the Ethics of Social Control', *International Journal of Ethics*, vol.40, no.1, pp.39-55; pp.44-48

Firstly (as it was with insights into the nature of brainwashing), my account may only offer this nuance to thinking within a certain area, albeit an area which, I suggest, covers much of the practical content of advertising ethics. As an ethics of change in 'moral' dispositions my account only speaks to the rights, wrongs, goods and bads of the shifting of moral dispositions. Given this, when applied to advertising ethics my account only speaks to some of the practical problem cases which comprise the discipline's subject matter; those which involve the influencing of dispositions which have a moral character. I suggest, however, that many of the dispositions influenced by advertising have such a character.

In §2.2 I identified moral character with connection to reasons which have a certain sort of overriding importance. For a disposition to be 'moral' it must be such that it at least partly determines behaviours for which there usually exist overriding 'moral' reasons for or against execution. Given this, I suggest, it should be clear that many of the dispositions targeted by advertisements have a moral character. This may be because the dispositions targeted cause behaviours which are themselves vicious or virtuous – such as the purchasing of morally problematic goods (for instance, goods whose production involves large negative externalities) or goods which somehow do good (by, say, bringing great joy to their users). It may also or instead be because the dispositions induced by the influences in question impose opportunity costs with a moral character, say by diverting income you'd otherwise use for some morally important project. It may also be because the good an advert influences you towards purchasing is adversarial and the loss of the good others encounter on account of your purchasing imposes some morally significant cost, such as suffering on their part (imagine you buy up most of the local water, leaving everybody else thirsty).

I suggest that many of the dispositional changes intended to result from the application of advertising techniques may be said to have some moral character on account of at least one of these species of entanglement. This is to say nothing of other dispositional changes which may be caused by advertising unintentionally and which may themselves have moral characters (such as, say, discriminatory tendencies encouraged by stereotyping in adverts[430]).

If the dispositional changes caused by advertising often acquire moral characters in such ways then it follows that the influences behind them, those employed within real-world advertising efforts, may often appropriately be judged by the standards of my account. The systematic exception to this comprehensiveness (amongst effective adverts) is the set of cases where the caused dispositional changes happen to be very unimportant, lacking in any moral character. Such cases must evade the moral entanglements I've suggested and not have morally significant unintended effects on dispositions. I suggest adverts which only change unimportant characteristics of existing dispositions, such as changing one from being a consumer of Product X made by more-or-less identical Company A to being a consumer of Product X made by more-or-less identical Company B[431], may escape my account's coverage in this way. Adverts that simply fail as adverts by triggering no dispositional changes are similarly exempt.

Amongst those adverts not exempted in such ways, however, the rationalizability requirement applies. This being the case, such adverts (whatever their form) ought to influence any impacted moral dispositions only by showing influenced agents reasons for the dispositional changes sought. If

[430] S.L.Grau and Y.C.Zotos (2016), 'Gender Stereotypes in Advertising: A Review of Current Research', *International Journal of Advertising*, vol.35, no.5, pp.761-770; pp.762-763

[431] Characteristic of advertising in a mature oligopoly, say, wherein competition by differentiation has ceased; *see* L.Severová, L.Kopecká, R.Svoboda and J.Brčák (2011), 'Oligopoly Competition in the Market with Food Products', *Agricultural Economics (Praha)*, vol.57, no.12, pp.580-588; pp.581-582

you want to sell somebody shoes and said somebody becoming disposed to buy shoes from you involves changes in moral dispositions (for some reason), you ought to show him why he should buy the shoes you're selling. Show him (say) how comfortable they are, why their manufacturing didn't damage the environment or how they themselves look. Don't (say) put them on a celebrity he likes, associate them with an irrelevant but pleasant song or, for that matter, show him things about them which aren't true and hence can't be reasons for anything. The reasons you show to your intended customer needn't be sufficient or have moral characters – 'relevance' doesn't demand anything so grand (though it doesn't preclude such things) – but they must be reasons, and they must be clearly presented or represented to those you're trying to influence, and they must cause the full influential effect of your advertising. An advertising project that meets these standards is an ethical one at least given the ethics distinctive to influencing moral dispositions. An advertising project which doesn't meet these standards and effects at least some moral dispositions necessarily harms its subjects insofar as it limits the extent to which their behaviour is determined by those things that ought to determine it.

The same may be said about the subclass of advertising known as propaganda. The term 'propaganda' carries strong connotations of dishonesty and manipulativeness, but I suggest that it's best defined by its objectives as a species of advertising which attempts influencing of certain paradigmatically moral dispositions, such as agents' political tendencies. This being the case, and 'propaganda' being understood as involving the influencing of dispositions instead of or in addition to the peoples' pictures of the world, propaganda by its nature is morally entangled and apt for judgement by the rationalizability requirement. Thus, the projects of the propagandist are bound by norms demanding influencing only by showings of reasons.

Operationalising these standards is no easy task. Given the considerations raised in Chapter 4, difficulties in the identification of reasons and their weighting present significant – if soluble – problems with applying any of the insights offered by my account. Mercifully, at least with respect to advertising, some of these problems can be dispensed with relatively easily. Selling people something – which is usually your ultimate goal in advertising a product, company or brand – is often not going to be important enough to justify alienating people from the reasons that ought to determine their behaviour (never mind damaging their autonomy or interpersonal freedom, as I've argued unrationalizable influences may).

Propaganda cases are more problematic insofar as they systematically involve intended moral-dispositional changes which impact important behaviours and thus systematically carry consequences which may be sufficient to trump concerns about the right and wrong ways of influencing.

Such trumping must however be proved case by case and, as I noted in §4.6, must contend with the fact that the very reasons that may be used to justify such trumping themselves may furnish influencers with the means to influence legitimately. If, say, one has strong reason to use the power of association in order to inspire someone to adopt certain dispositions on account of the fact that they are life-saving, one necessarily has access to a powerful tool of legitimate influencing in the form of the fact that the dispositions sought are life-saving. If this fact could be made vivid for proposed influence-subjects it might be possible to influence them as desired without needing to resort to illegitimate methods like leveraging associations.

Ethical propagandising thus requires both a cause sufficient to justify whatever alienation from reasons one inflicts and there being no alternative more legitimate means available by which to

effect dispositions, even though there's likely to be at least one such means available given this sufficiency.

Any application of my views to the practice or regulation of advertising must also contend with the fact that advertising often targets diverse sets of people, some of whom may have different reasons for action and will thus be such that different things are relevant for them. You might enjoy the taste of chocolate, and thus be such that an advert which represents chocolate-eating as enjoyable may legitimately influence you towards buying it, while I might hate the taste of chocolate and thus be such that the same advert cannot legitimately influence my behaviour in the same way using the same representations (presuming, of course, that in the case at hand chocolate-eating somehow has some moral character and that the enjoyable taste of chocolate constitutes a reason to buy chocolate).

Given this diversity amongst potential subjects of even individual advertising projects, I suggest that any workable ethics of advertising which embeds the rationalizability requirement must do so by tolerating at least some generalisation and associated harm. One must be willing to tolerate advertisers making reasonable assumptions about an advert's audience and what reasons they have and accept the advertiser structuring his efforts accordingly, so long as he grounds their full influential content in these assumed reasons. In doing so one must be willing to tolerate (say) relatively rare chocolate-haters sometimes being presented with images that show the enjoyability of consuming chocolate, even though these images on account of these agents' chocolate-hating cannot play any proper role in their dispositional development. One must even be willing to tolerate such chocolate-haters sometimes being successfully influenced to eat chocolate by representations of how pleasant eating chocolate is, representations which cannot legitimately influence such agents. One must be willing to tolerate such things, at least, if one isn't to reject all advertising save that which is tailored to specific agents or else owes its effectiveness to its presentation or representation of universal reasons (if there are such things).

Such 'problems of generalisation' and consequent tolerance requirements often arise from the technical conditions under which advertising takes place. An ethical advertiser – who doesn't want to influence illegitimately – may still very easily find herself somewhat forced to do so (at least insofar as she wants to stay in advertising) simply because the only means she has to advertise are bad at discriminating amongst who they reach. If your only practical way of advertising is a billboard by a main road, you aren't in any position to choose who you reach so much as whether you reach nobody or a moderately random set of people.

Targeting is the salvation from this technical problem, short of radically limiting the influential content of adverts to representations of truly universal reasons or else somehow eliminating mass advertising and everything like it from society. Targeting can be achieved in many ways. While it's strongly associated with various (often nowadays digitally-enabled) practices which work by picking out which agents encounter which adverts[432], it can also work by other methods such as the use of designs or locations intended to attract the attention of specific audience segments.

Targeting, however it works, can be used to select who is influenced by your advertising efforts and in doing so meet the demands of the rationalizability requirement. By making sure that your adverts only influence those who have the reasons for action your adverts represent, you can ensure that your adverts only influence in legitimate ways. It's noteworthy that those advertisers that stress

[432] *see for example* F.Caruso, G.Giuffrida and C.Zarba (2015), 'Heuristic Bayesian Targeting of Banner Advertising', *Optimisation and Engineering*, vol.6, no.1, pp.247-257; pp.252-253

their targeting capabilities, when defending these capabilities from a hostile public (typically – in a contemporary context – one concerned about privacy[433]), often invoke this relatively noble purpose of targeting, arguing that it facilitates the delivery of more 'relevant' materials and in doing so serves the interests of advertising-subjects. My account gives meaning to these defences: in increasing the relevance of the content of advertisements, in my view, one increases the extent to which any resulting influence maintains valuable connections between subjects' reasons and dispositions they ought to effect.

Unfortunately, however, this 'noble purpose' (sometimes used to justify the proliferation of targeting systems in the contemporary advertising industry[434]) fails appallingly to capture the whole effect of said systems on said industry or the rationale behind their contemporary use. Targeting doesn't just allow for more ethical influencing in advertising; it can be used to bolster the effectiveness of advertising including the effectiveness of unethical forms of influencing used in advertising.

Consider that different agents may be such that they have differing vulnerabilities to various forms of influencing, and that these differing vulnerabilities can be entirely other than justified contextual or subjective divergences in reason-responsiveness. I might be a jazz fan, for example, and thus be such that I'll be more positively inclined towards any product or message displayed in association with jazz music[435]. I might be prone to anxiety, and thus relatively more responsive to messages which evoke fear, independent from said messages' actual content[436]. I might be somebody who consumes media in a certain pattern, and thus be such that a system which is able to detect this pattern and make sure that it presents a single advertising message at multiple junctures in this pattern will be able to accentuate the influential effect of said message by force of repetition[437].

Targeting may facilitate the exploitation of traits like these to strengthen the influential power of adverts, but it's hard to argue that such exploitation could constitute reason representation or presentation. In what ordinary context, say, could the fact that your favourite piece of music was played during an advert for some shoes show any reason to buy said shoes? In what context could the repetition of an advertising message add anything to its rational content?

Advertising, methodologically dodgy or not, costs money and the primary advantage of targeting for the profit-motivated advertiser is precisely that it allows them to deliver their material only to those upon whom it'll have the greatest influential effect. In essence, targeting allows for improvements in the effectiveness of advertising in terms of how much change in behaviour one's money can buy. This is the case whether the underlying methods used in the advertising in question involve showing reasons or not. Thus, from the (partial) perspective of my account, targeting is best understood as

---

[433] J.Hinds, E.J.Williams and A.N.Joinson (2020), '"It Wouldn't Happen to Me": Privacy Concerns and Perspectives following the Cambridge Analytica Scandal', *International Journal of Human-Computer Studies*, vol.19, no.8; p.3, 6-7 (issue pagination unavailable)

[434] *see for example* D.Kirkpatrick (2016), 'Study: 71% of Consumers Prefer Personalised Ads', *Marketing Dive*, available at: https://www.marketingdive.com/news/study-71-of-consumers-prefer-personalized-ads/418831/ [webpage] [accessed 24/1/2021]

[435] W.Raja, S.Anand and D.Allan (2020), 'How Ad Music Attitude-Based Customer Segmentation can help Advertisers', *Journal of International Consumer Marketing*, vol.32, no.5, pp.383-399; pp.383-384

[436] A.Nix (2016), 'The Power of Big Data and Psychographics' [presentation], 2016 Concordia Annual Summit, New York, available at https://www.youtube.com/watch?v=n8Dd5aVXLCc [video] [accessed 2/8/2018]

[437] S.Schmidt and M.Eisend (2015), 'Advertising Repetition: A Meta-Analysis on Effective Frequency in Advertising', *Journal of Advertising*, vol.44, no.4, pp.415-428; p.416

something which raises the ethical stakes of advertising; it may facilitate relevance but also the effectiveness of the irrelevant.

Given the massive growth in the targeting capacity of the advertising industry in the past twenty years, with the largest companies in the contemporary business dedicated to facilitating this capacity by exploiting digitalisation, this fact about targeting should cause concern. The problem here isn't novel, of course; an old-fashioned billboard may exert illegitimate influence. What's novel, though, is the sheer and growing capacity for targeting which comes with life moving online and the consequent potential for more effective illegitimate influencing. The contemporary advertiser might know (or control a computer that, in some sense, 'knows') that you just broke up with your partner, love jazz music or support liberal politicians. She's thus enabled to devise methods to change your dispositions which are structured by these facts, methods which may exploit psychological vulnerabilities implied by these facts (as uncovered, perhaps, by sophisticated inductive algorithms[438]) in place of providing relevant reasons.

I think the use of such techniques and the cultivation of interpersonal power which – as I argued in §5.11 – comes with them explains some of the public discontent which has accompanied the entrenchment of the internet as an advertising tool. This discontent in turn has occasionally spilled over into moral panic, notably during the Cambridge Analytica affair[439]. Such panics have been characterised by demands for increased privacy and have been intermittently answered at a policy level in such terms, with tightening of legal control over data flows. If I'm right, though, we shouldn't understand such demands for privacy as demands for privacy as an end-in-itself (as with our demands for privacy regarding how we look when we use the bathroom). Rather we should understand these demands as demands for privacy to be used as a safeguard against targeted advertising which threatens the sorts of interpersonal freedom and reason-responding agenthood we aspire to. We object to advertisers knowing too much about us not because we just don't want them to know certain things but rather because we don't want them to be able to do certain things to us – that is, exert certain sorts of influence.

Given this, it follows that there are limits to what increasing privacy can achieve with respect to satisfying the needs that lead people to demand it. If one makes targeting harder one risks preventing relevance-facilitating targeting practices that facilitate a valuable sort of agenthood as well as relevance-disregarding practices which deplete such agenthood. Thus, in my opinion, attempts to correct current and future developments in advertising ought to focus more on 'fixing the ads'; purging the advertising industry of methodologies which work by leveraging influences other than relevant reasons. There have been relatively few consistent attempts to do this in the past. Subliminal techniques have been legislated against[440] but we've normalised other techniques such as association or repetition despite them being no more dependent on relevant content. In the UK, at least, using threats to scare people into purchasing things is banned[441], but fear is hardly the

[438] A.Nix (2016), 'The Power of Big Data and Psychographics' [presentation], 2016 Concordia Annual Summit, New York, available at https://www.youtube.com/watch?v=n8Dd5aVXLCc [video] [accessed 2/8/2018]
[439] J.Hinds, E.J.Williams and A.N.Joinson (2020), '"It Wouldn't Happen to Me": Privacy Concerns and Perspectives following the Cambridge Analytica Scandal', *International Journal of Human-Computer Studies*, vol.19, no.8; p.1-2 (issue pagination unavailable)
[440] M.R.Nelson (2008), 'The Hidden Persuaders: Then and Now', *Journal of Advertising*, vol.37, no.1, pp113-126; p.117
[441] The Consumer Protection from Unfair Trading Regulations 2008, SI 2008/1277, part 2, available at http://www.legislation.gov.uk/uksi/2008/1277/pdfs/uksi_20081277_en.pdf [pdf] [accessed 24/1/2021]; p.4

only problematic means of influencing available to the advertiser. Future attempts to regulate away problematic techniques in advertising must improve on such starting points.

If we're able to manage this task, somehow, we might be able to save what's good about targeting – its capacity to direct advertising towards those who have the reasons for action it represents – while rendering it less of a threat to reason responder-hood. How exactly this may be accomplished is ultimately a political question about which I cannot comment here. Such reforms also may or may not combat the tendency to polarise that targeting can have[442], depending to what extent this polarisation in practice and necessarily requires the leveraging of irrelevancies.

To see why such reform is needed, consider the future that is offered by maintaining the status quo.

Suppose that the methods of the advertising industry continue to be shaped by profit motives and perhaps norms precluding outright lying, and that the future holds ever-increasing advertising-driven penetration of data collection and digital mediation into human experience. In such a future, I suggest, if we don't somehow ensure that relevant reasons occupy a privileged place in the influential armoury of the advertising industry, we'll find ourselves progressively finding our behaviour detached from the world we inhabit and the reasons for action that exist within it.

We'll go on acting, of course, but the causal history of our actions will trace back to dispositions introduced and maintained by influences unconnected from whatever reasons are relevant to our actions, influences which instead pull whatever psychological levers they find most convenient for their purposes. Life in such a world will involve being a reason-responder only as much as is convenient to the infrastructure of influences that pervades life. We'll become beings whose actions are conditioned not by their conditions but ultimately by the effective 'will' of the influential powers we are subject to and whoever determines said will. This, I think, will necessarily involve sacrificing something very important; the possibility of a life lived in response to the realities around one and the reasons said realities present. It'll be like a life in an experience machine, alienated from what matters for the sake of (likely somebody else's) convenience. It'll also be a life less interpersonally free or, in any appealing understanding, autonomous.

(6.5) Biomedical Moral Enhancement

What, then, does my account say about evaluating various means of enhancing morals? As I noted in §3.7, Nozick suggested that the demand for a proper connection with reality he highlighted may be used to explain why a key part of debates over the use of psychoactives contests whether such use severs one from, or somehow connects one better with, reality[443]. Following the rationalizability requirement, an analogous way of thinking may be applied to moral enhancements, including biomedical moral enhancements.

In §3.11 I highlighted a debate over the administration of oxytocin as a method for enhancing morals. Some within this debate construe oxytocin as an 'empathy-enhancer', a compound which somehow heightens awareness of the feelings of others and in doing so makes people more aware of things which plausibly constitute reasons that should impact their treatment of others[444]. Others within this debate argue oxytocin is more like a 'bonding-enhancer' which triggers tendencies to

---

[442] S.Bradshaw (2019), 'Disinformation Optimised: Gaming Search Engine Algorithms to Amplify Junk News', *Internet Policy Review*, vol.8, is.4, pp.1-24; pp.2-3

[443] R.Nozick (1972), *Anarchy, State and Utopia*, New York, Basic Books; pp.43-45

[444] N.Levy, T.Douglas, G.Kahane, S.Terbeck, P.J.Cowen, M.Hewstone and J.Savulescu (2014), 'Are You Morally Modified? The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-171; p.117

treat others one feels bonded with (by ties of family, tribe, even race) better[445]. I suggested that, if my account holds, this debate is relevant to the legitimacy of oxytocin as an influence on moral dispositions.

This debate is relevant because empathy is something integrated with perception and, when it's working well, something which can plausibly help us to perceive reasons to change our behaviour. If oxytocin enhances empathy, then it plausibly works by making us aware of things (others' feelings) that constitute reasons to change our moral dispositions (in the way of a good influence enabler, described in §4.5), in the same way as the revelation of some relevant fact might. This is acceptable on my account. Nothing wrong with seeing the world more aright than one used to and being moved by one's enhanced perspective. If oxytocin intensifies behavioural responses to bondedness, on the other hand, then it's hard to construct its action in terms of the showing of plausibly relevant reasons (directly or through some enablement). One's feelings of bondedness provide a less plausible reason for treating somebody well than their feelings, and if oxytocin works by boosting the effect of feelings of bondedness on behaviour rather than modifying feelings such that they better show reasons then it cannot be construed as showing (or causing the showing of) anything relevant.

There are many potential complexities to this case which, depending on the empirical matter of exactly *how* oxytocin works, may change the results of my analysis as applied to it. If, for instance, empathy is understood not as a sort of sense and rather as, say, a motivational module with no representational or presentational capacity, then oxytocin administration for moral enhancement could not be legitimated in the way I suggest. Similarly, if oxytocin is understood as directing the agent's attention to their bonds with others in some way which can successfully show these bonds (or enable such showing), and these bonds are understood as being potential reasons for the sorts of treatment oxytocin promotes, it's possible to legitimate moral enhancement by oxytocin administration on my account.

This case shows us two important things about the application of my account to the evaluation of biomedical moral enhancements.

First, the case shows that any attempts to accomplish such evaluation must depend upon nuanced empirical psychological knowledge of the enhancement in question. You must know whether the dispositional effects of the enhancement are achieved by some sort of showing of something to the agent, somehow. You must also know to what extent the effect of this showing on the agent is responsible for changes in their moral dispositions. You must lastly know what contents are shown in this showing, so that they may be checked for relevance.

Such detailed knowledge needn't be available or easy to acquire and where it's missing, or suspect, any evaluation of influencing based upon such knowledge must be tentative. Advancing this knowledge base and articulating methods for how we might evaluate influences on morals given gaps in this knowledge are problems beyond the scope of this thesis; the former demands empirical research while the latter requires the application of decision theory. I do suggest, however, that it's at least incumbent on anyone arguing for the application of some means of biomedical moral enhancement to describe precisely how said means works and thereby show to what extent (if any) they satisfy the rationalizability requirement. Absent such information, given the rationalizability requirement, fulsome evaluation of a method of biomedical moral enhancement is impossible.

---

[445] *ibid*. p.118

As evaluations of methods of moral bioenhancement depend upon empirical detail in this way, given my account, it's impossible to use my claims to ground universal statements about the acceptability or merits of biomedical moral enhancement. Different methods of biomedical moral enhancement can work in different ways and as the way in which such methods work matters (on my account) there can be no attempt to evaluate all such methods together.

Second, the oxytocin case shows that given the rationalizability requirement two matters are salient in evaluating moral bioenhancements: (1) does the bioenhancement work by showing anything and (2) does what the bioehancement shows (or causes the showing of) constitute a reason for the dispositional changes it generates? These are the same two matters that are always salient in the application of my thesis but the first of these carries a special significance with respect to moral bioenhancement.

As I argued in §4.5 bioenhancements – like physiological interventions in general – which work directly on human bodies are capable of changing dispositions in unusually direct ways. It is possible, say, for such an intervention to change one's physiology so as to dispose one towards entirely new behaviours, without mediating this effect through one's perception and thus even possibly showing reasons to one (for one must perceive something in some sense for it to be successfully presented or represented to one).

Addictive chemicals which build physiological dependency provide the most familiar example of this. Irrespective of what you experience if you consume alcohol excessively for long enough you will find yourself disposed to keep drinking it, a disposition enforced by unpleasant and hazardous withdrawal effects should you try to drink less[446]. Getting into this state is not a matter of whether – say – you enjoy drinking or not but a matter of your physiology adapting to your alcohol intake such that it becomes difficult for you to drink less[447]. Acquiring the dispositions of an alcoholic is thus (at least typically) not a matter of perception or one's reactions to one's perceptions; it's a matter of what alcohol does to your body.

For different purposes, and with different methods, many programs of biomedical moral enhancement work in analogous ways. Consider, for instance, the early 20th century practice of lobotomy[448]. In this tradition psychosurgeries were used to effect changes in dispositions – they would, for example, be used to render violent patients more docile[449]. These procedures patently did not achieve these effects by showing patients relevant reasons; they rather involved directly destroying or disrupting neurophysiology so as to cause dispositional change[450]. Such bioenhancements, by their nature, must be declared pro tanto illegitimate on my account, at least insofar as they alter moral dispositions. The legitimacy of any method of moral-dispositional change depends upon reasons shown to the changed subject being causally responsible for the subject's moral-dispositional change(s). Thus, any influence on moral dispositions which fails to show anything to its subjects – or at least anything causally responsible for their moral-dispositional changes – will

---

[446] A.McKeon, M.A.Frye and N.Delanty (2008), 'The Alcohol Withdrawal Syndrome', *Journal of Neurology, Neurosurgery and Psychiatry*, vol.79, no.8, pp.854-862; p.854

[447] *ibid.* pp.854-855

[448] *see* B.M.Collins and H.J.Stam (2015), 'Freemans Transorbital Lobotomy as an Anomaly: A Material Culture Examination of Surgical Instruments and Operative Spaces', *History of Psychology*, vol.18, no.2, pp119-131; pp.120-125

[449] R.Cooper (2014), 'On Deciding to Have a Lobotomy', *Medicine, Health Care and Philosophy*, vol.17, no.1, pp.143-154; p.145

[450] *ibid.* pp.144-145

be declared pro-tanto illegitimate on my account. Such interventions thus, in my view, must be somehow shown (case-by-case) to be lesser evils, or else opposed.

This illegitimacy, furthermore, is present (more or less unfortunately) whether-or-not the biomedical intervention in question is intended as a moral bioenhancement, for the harm such methods do in alienating agents from the reasons that ought to shape their behaviour doesn't vary with intent. Thus, for instance, one must show the use of a recreational drug with direct effects on moral dispositions (analogous to my alcohol example, say) to be a somehow lesser evil, or else oppose said use on ethical grounds. Accidental moral-dispositional influences which fail to work through the showing of reasons, such as the effects of random poisonings or blows to the head, may be evaluated along the same lines. Whenever such things influence your moral dispositions, however they do this (and with whatever effects) at least part of the harm they do to you, part of the tragedy of the accident, will consist in your alienation from reasons that ought to determine your behaviour.

Biomedical moral enhancements, then – and indeed other physiological interventions which change moral dispositions – provide an important context for the application of the rationalizability requirement to practical ethical problems. Such application, which will always be contingent upon the actual workings of such interventions (and thus empirical knowledge), will involve a two-stage evaluative process of testing for representational or presentational content and then testing the relevance of said content. Some methods of moral enhancement, I suggest, will run afoul of the first of these stages due to their nature as direct interventions in the physiology determining moral dispositions. I argue we have some pro tanto reason to reject the use of these methods.

(6.6) Nudging

Another debate which could benefit from the application of ideas developed in this thesis concerns the ethical use of 'nudging'.

In the form popularised by Sunstein, to 'nudge' is to exert influence over behaviour (moral or non-moral) through the application of some sort of method that avoids coercion or strong incentivisation and leaves open the possibility of choosing otherwise, but nonetheless increases the probability that the nudged agent will make some specified choice[451]. Amongst the set of all nudges one may find things such as the exploitation of cognitive biases (such as, say, establishing preferred options as defaults), education campaigns intended to acquaint nudged agents with clearly superior options, curated simplification of options, and myriad uses of the armoury of contemporary advertising[452].

It's little wonder that such a diverse set of interventions attracts ethical debate, not least about the benefits of lumping all such things together as 'nudges'. It may be that all nudges share some common moral advantage, say inasmuch as they all preserve a certain formal capacity for free choice[453], but even if this is the case clearly there's more evaluative detail necessary to properly evaluate the use of such a diverse set of influences. This is where I suggest my account may help.

Sunstein, in carving up nudges into more and less acceptable forms, broadly distinguishes educative from non-educative nudges[454], arguing on consequentialist grounds that educative nudges are to be preferred 'sometimes' but cautioning that in some circumstances 'default rules' are preferable[455]. At

---

[451] C.Sunstein (2014), 'Nudging: A Very Short Guide', *Journal of Consumer Policy*, vol.37, no.4, pp.583-588; pp.583-584
[452] *ibid.* pp.585-587
[453] *ibid*. p.584
[454] C.Sunstein (2016), *The Ethics of Influence*, New York, Cambridge University Press; p.32
[455] *ibid.* pp.33-34

the same time[456] Sunstein notes that people seem to prefer educative to non-educative nudges[457]. This preference in my view plausibly responds to the foundational claim of my account; that there's some value in having one's behaviour determined by relevant reasons rather than other factors. Sunstein's 'educative' nudges, which work by showing agents realities which include reasons for behavioural change, are – I suggest – preferred precisely because their workings respect this value.

These same 'educative' nudges, however, are also understood by Sunstein as working more-or-less cognitively[458], and are thus distinguished from other nudges which, say, offer graphic warnings and work through sheer emotional punch[459]. These emotive nudges, in turn, are bundled in with various exploitations of behavioural biases as problematically manipulative (though sometimes justifiable), a perspective which ignores the fact that graphic warnings are sometimes true and owe their motivational power to the ugly realities they show, quite unlike mere exploitations of behavioural biases. For example, a graphic image of a lung damaged by smoking may put one off smoking mostly by triggering certain emotional responses, but despite this it no less shows a reason to change behaviour in the form of a representation of what smoking often does to one's lungs. In such situations, I suggest, any 'manipulation' which occurs is of a rationalizable and pro tanto legitimate sort[460]. As I argued in §5.5 it's legitimate for emotions to effect moral dispositions provided these emotions properly mediate between showings of reasons and dispositions. Sunstein nonetheless commits to evaluating all emotive nudges alongside influences bereft of presentational or representational content, seeing all of them as irrational 'system 1' (in Kahneman's sense[461]) methodologies[462] rather than as a set with varied evaluative texture determined by relations with relevant reasons. Inasmuch as he does this Sunstein tacitly accepts the view that there's something special about beliefs that makes them especially legitimate moral influences. Such a view, I argued in Chapter 5, amounts to an objectionable moral fetishism.

I discuss Sunstein here for he's the most significant author discussing the ethics of nudging, but others within the same literature – including some of Sunstein's critics – share the same problematic perspective on this point[463]. For or against nudging, there's a temptation to divide nudges into the rational and irrational using their relations to cognition and in doing so leave out the important nuance of the rationalizable but belief-unmediated. This is a pattern of debate which maps closely on to debates about the acceptability of advertising methods discussed in §6.4 and which may be analysed, as per my account, in the same way.

There is, however, one feature of nudging which distinguishes it from advertising in a way that matters from the perspective of my account. The advertiser (at least if she be only an advertiser) might control messaging about a product but she doesn't control the product itself. She doesn't, say, price what she's selling, or decide where or how it's sold; she simply tries to sell it. The 'nudger' on the other hand paradigmatically has access to a wider field of action, for he's empowered to

---

[456] *ibid*. pp.32-33
[457] *ibid.* pp.32-33
[458] *ibid*. p.33
[459] C.Sunstein (2014), 'Nudging: A Very Short Guide', *Journal of Consumer Policy*, vol.37, no.4, pp.583-588; p.586
[460] At least assuming that the sort of modulation of the content of the influence highlighted by Velleman's 'Problem of Representation' (*see* J.Velleman (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371; pp.365-366), discussed in §3.10, is not too great a factor.
[461] *see generally* D.Kahneman (2011), *Thinking Fast and Slow*, Farrar, Straus and Giroux, New York
[462] C.Sunstein (2016), *The Ethics of Influence*, New York, Cambridge University Press; p.34
[463] B.Engelen and T.Nys (2020), 'Nudging and Autonomy', *Review of Philosophy and Psychology*, vol.11, no.1, pp.137-156; pp.137-140

influence broader features of the 'choice architecture' encountered by those he seeks to influence. These powers may include control of messaging, but also control over the mechanisms by which choices are expressed and, to an extent, the characters of the choices being made.

For example, a nudger might seek to encourage charity by making it easier to undertake, say by increasing the number of donation points amidst the nudged population. Such an adjustment could plausibly impact moral dispositions towards charitability, as (say) agents who start giving more because they more frequently encounter opportunities to do so form habits which encode stronger dispositions towards charitability. In a case like this influencing is accomplished not by some sort of message or piece or media, as with advertising, but rather by changes in the world influence-subjects encounter introduced by the nudger.

These changes, in turn, may or may not constitute reasons in favour of the dispositional changes they encourage.

The number of donation points for a charitable cause, say, plausibly doesn't constitute a reason in favour of donating to said cause, whether-or-not it encourages donation or tendencies to donate. Provided this is the case given my position it follows that increasing the number of donation points for a charity is not a legitimate tactic to increase dispositions towards charitability as per the rationalizability requirement (though it may of course still sometimes be justified, all-things-considered). This is the case because the effective influence in such a case – the frequency with which influenced subjects encounter opportunities to donate – fails to show a reason to donate.

Consider though a different case where our putative nudger, instead of increasing the number of donation points, takes steps to modify donation procedures (something the nudger is permitted to do, while remaining a nudger, provided said modifications don't strongly alter incentives or introduce coercion[464]). For example, suppose the nudger effectively makes donation cheaper (in time terms), by making it easier to accomplish – say by replacing an onerous form with a simple donation box. Suppose further that this change influences dispositions towards donating, say by making donation initially more compelling and eventually more habitual. In such a case, I suggest, the influence generated by the nudger's modification is plausibly pro tanto legitimate because the 'nudge' in question is composed of the introduction of additional reason which is presented to the nudged party (who percieves he no longer needs to fill in an onerous form to donate, and hence that he can now do as much good with less effort).

That this reason was added by the nudge itself is immaterial: as the influential effect is achieved by this relevant reason being presented, the rationalizability requirement is satisfied. This capacity to not merely show but also add and show reasons differentiates the nudger from the advertiser or propagandist, who may only show things (be they reasons or otherwise). It also opens the door to an additional general method of legitimate influencing: creating and presenting influence-subjects with additional reason to change their dispositions as you want them to.

This sort of influencing may or may not involve 'nudging'. 'Nudges', by definition, cannot alter the incentives of influenced agents *much*[465], but there are other methods of influencing behaviour which certainly can. If one wants somebody to act a certain way, a time-honoured way to achieve this is to pay him to do so. Similarly, another time-honoured way to achieve the same result is to threaten to fine, imprison or otherwise harm said person for acting otherwise. Such practices influence dispositions, insofar as they do so, by introducing new (and shown to the influenced party) reasons

[464] C.Sunstein (2016), *The Ethics of Influence*, New York, Cambridge University Press; p.5
[465] *ibid.* p.21

in favour of acting the way the influencer wants one to. For example: additional to other reasons in favour of, say, not consuming polluting products, you can now buy less polluting alternatives for a price forced down by subsidies and thus have more reason to do so. Additional to other reasons against, say, larceny, one now knows one will likely go to jail if one adopts larcenous dispositions and thus has more reason not to do so. Provided influencing of moral dispositions in cases where influencers are modifying the realities that influence agents in such ways proceeds only by agents responding to the reasons contained within these modified realities (that they perceive), dispositional changes in such cases will retain the virtue of rationalizability.

This isn't to say that altering behaviour by altering the set of relevant reasons can't, say, amount to problematic coercion. Giving someone a reason to Y by threatening 'Y or I'll kill you' is an awful thing to do, even if the threatened agents' response is rationalizable (it's a proper response to the reason shown to them by the threat). This also isn't to say that in such cases illegitimate influencing cannot occur. An agent told to Y on pain of violence may be paralysed by fear out-of-proportion to an ordinary fear of violence and thus come to Y on account of something other than a response to the reason represented in the threat (furthermore, ethics of influencing aside, threatening people remains coercive and bad). My point here is only that my standards for changing moral dispositions remain constant even when the reasons relevant to the change in question are themselves dynamic and under somebody's control. Nudges – being a class of influences that includes some reason-making as well as reason-showing and the exploitation of influential irrelevancies – are such that they cannot be fully evaluated without this point being made. Given this point, though, I suggest that nudges may be evaluated by dividing them into three subsets.

One of these subsets will be composed of the sort of nudges teased out by the complexity noted here; nudges which work by changing the balance of reasons in favour of or against enacting a behaviour, and thereby influencing dispositions towards or against enacting it. Such nudges, provided they influence by showing to influenced agents the changes to the set of relevant reasons they have introduced (perhaps by leaving agents to discover the changes themselves), will have the virtue of rationalizability.

The next of these three subsets will include nudges which do not change the balance of reasons and instead work by showing existing reasons to influenced agents, perhaps by improving the vivacity of such showings. This subset will include those nudges which Sunstein calls 'educative[466]', but also some nudges which work by directly inspiring dispositional change by showing relevant reasons without mediating beliefs between these influences and dispositional changes. This subset is highly analogous to the set of legitimate advertisements described in §6.4, perhaps unsurprisingly as the literature on nudging draws heavily on advertising research and concepts[467]. Like their analogues, these nudges are legitimate influences on moral dispositions by the lights of the rationalizability requirement.

Analogous to the earlier described set of illegitimate advertisements, in turn, is the third subset of nudges relevant to the ethics distinctive to moral-dispositional change. This subset includes nudges which achieve their effects without showing relevant reasons to agents. They may fail this content-test, in the way of some adverts, because they show only irrelevant things to influenced agents. They may also, like certain biomedical moral enhancements, fail this content-test by failing to show

[466] *ibid*. pp.32-34

[467] *see for example* J.Kim, M.Giroux, H.Gonzalez-Jimenez, S.Jang, S.Kim, J.Park, J.E.Kim, J.C.Lee and Y.K.Choi (2020), 'Nudging to reduce the Percieved Threat of Coronavirus and Stockpiling Intention', *Journal of Advertising*, vol.49, no.5, pp.633-647

anything at all to agents. For example, the order in which the options on a questionnaire are listed plausibly represents nothing to those completing said questionnaire. True, people might carry a certain bias in favour of the first option listed in such cases, but it would be a brave interpreter who tried to give this tendency meaning beyond being a simple manifestation of laziness. Redesigning questionnaires which somehow effect morally significant behavioural dispositions (for example, ballot papers at elections) to put certain options first in any listings that appear is thus an illegitimate means of influencing said dispositions.

Like other influences, nudges may escape the ethical constraints I propose in cases where they influence non-moral dispositions, fail to influence any dispositions or influence agents extremely poor at reacting to reasons. They also might be justified, despite being illegitimate influences, in cases where they do sufficient good to outweigh the harm they do in alienating their subjects from the reasons that ought to influence their behaviour. Furthermore, like other influences, the relative strength of the ethical constraints that the rationalizability requirement generates with respect to nudges may be assessed according to the significance of the moral dispositions subjected to influence, the numbers influenced and the effectiveness of the influencing in question.

What practical significance do these claims have with respect to the business of nudging? The concept of nudging means to carve out a space for the direction of human behaviour which is purged of the sort of coercion abhorred by a certain liberal political sensibility[468]. It's meant to be a way of controlling people's behaviour without bossing them about, because 'bossing about' is some combination of unpopular, inefficient and wrong[469]. This tradition has never been naïve enough to think that the satisfaction of 'no bossing' requirements is all there is to the rights and wrongs of controlling behaviour; some behaviours are themselves good or bad, for one thing. It has however sometimes indulged the thought that it can (at least largely) sort out all the rights and wrongs of controlling behaviour *which pertain to influencing specifically*, by successfully ruling out (or at least relegating to 'lesser evil' cases) coercion[470].

If I'm right, though, there exists a class of influences which are bad and consequently sometimes wrong to use in virtue of their characters as influences which nonetheless needn't be coercive. This class is the class of unrationalizable influences which alienate agents from reasons relevant to their behaviour. In neglecting to consider this class, and the species of defect of influences *as influences* that unifies it, I suggest that extant work on nudging misses a significant nuance of how we ought to influence human behaviour. Given this, nudging – as presently defined – may in some sense conserve freedom but nonetheless sometimes leave one troublingly disconnected from the reasons relevant to one's conduct. Debate about nudging should benefit from eliminating this blind spot.

(6.7) Conclusions

I've argued in this thesis that changing moral dispositions has an ethics. This ethics, I've suggested, is grounded in the valued place that reality and the relevant reasons it contains have in the development of human behaviour. I've argued that there's an important sense in which we should distinguish between pro-tanto legitimate and illegitimate influences by attending to whether these influences show reasons to their subjects and whether this showing causes their effects. I've addressed challenges to this view of things from multiple quarters and shown how this view coheres

[468] A.Barton and T.Grüne-Yanoff (2015), 'From Libertarian paternalism to Nudging – and Beyond', *Review of Philosophy and Psychology*, vol.6, no.3, pp.342-359; pp.346-347

[469] C.Sunstein (2014), 'Nudging: A Very Short Guide', *Journal of Consumer Policy*, vol.37, no.4, pp.583-588; p.583

[470] C.Sunstein (2016), *The Ethics of Influence*, New York, Cambridge University Press; pp.200-202

with relevant conceptions of freedom. I've sought to integrate my account with broader ethics, shown how it may support clear claims despite background substantive disagreement, and then made some indication of what it has to say about a number of live and important debates. Why have I done all this? What purpose do I hope my insights to serve?

As I've emphasized throughout this text humanity is presently becoming aware of the conditionality of its own practical conscience. Stumblingly, imperfectly, inexorably, we're discovering the causal triggers – and levers – that shape our behaviour. This advancement in knowledge, I think, is more inclined to trigger fundamental changes to the human condition and experience than most others presently underway. While people of the past might have moved from passion to passion, or carefully plotted paths through life, in anarchic negotiation with a wild and inconstant environment, future lives may be the more-or-less predictable results of systems of influence devised by the people of today and tomorrow. With so much power spilling on to the table, after all, what are the chances that nobody will seize it?

The goal of this thesis, then, is to temper the consciences of whoever comes to seize this power. If we're to be – and I suspect in the future we might forever be more so – products of influencing, it's my hope that the ideas developed here might equip a few of the future's influencers to influence us as we should be influenced, and thereby preserve something valuable. To this end I have not spoken much about the ends of influencing. I've rather concentrated on the question of what things ought to influence us and I've proposed that, at least with regards to the morals we practice, it matters that these things be reasons.

I've used this perspective to illuminate many debates old and new about the proper place of influencing in human life. I've shown what it is to 'brainwash', even in a good cause, challenged purely instrumentalist ways of relating to one's moral dispositions and argued that influencing only with reasons helps preserve important freedoms. In doing these things I hope to have shown in some detail not merely how one should go about influencing in this-or-that circumstance, but to have offered a method by which we may evaluate influencings of (at least) our moral dispositions. It's my hope that this method – or some development of it – might help to preserve a valuable part of the human experience. A part which requires that our dispositions be shaped by the reasons we have for action.

I posed a question at the beginning of this thesis. I asked, if morals may change and we may control this change, how should we exercise this control? If I'm right, then there's a sense in which the answer to this question is that one shouldn't try to exercise such control, at least without some overriding justification and where the morals in question are those that are practised. Relevant reasons are the only legitimate influencers of moral dispositions, all one may ethically do (absent overriding justification) is make oneself a medium for their legitimate influence. This doesn't deprive one of choices, choices about which relevant realities to go out of one's way to show, practical choices about how to render such realities vivid; these remain ours. What isn't in anybody's ethical gift, though, (absent overriding justification) is going beyond relevant reasons in influencing; adding compunction where there's none through (say) association, repetition or rhetoric, or moulding dispositions without showing anything to the moulded agent (who may even be oneself). The influencer of moral dispositions is thus, in my view, entitled to all and only the power they may wring from the reasons there are to make the changes he or she intends. They are, though, given only the ethics distinctive to influencing moral dispositions, entitled to add nothing more. All other things being equal one's practised morals should be products of one's reasons and nothing else.

Bibliography

Aboodi, R. (2017), 'One Thought too Few: Where De Dicto Moral Motivation is Necessary', *Ethical Theory and Moral Practice*, vol.20, no.2, pp.223-227

Adams, R.M. (1995), 'Moral Faith', *The Journal of Philosophy*, vol.92, no.2, pp.75-95

Aizenkot, D. (2020), 'A Quantitative and Qualitative Approach to Analysing Cyberbullying in Classmates' Whatsapp Groups' in Nguyen, D., Dekker, I. and Nguyen, S. (2020) [eds.], *Understanding Media and Society in the Age of Digitalisation*, Cham, Palgrave Macmillan, pp.185-208

Allison, H. (1990), *Kant's Theory of Freedom*, Cambridge, Cambridge University Press

Arpaly, N. (2002), *Unprincipled Virtue: An Inquiry into Moral Agency*, New York, Oxford University Press

Arpaly, N. (2002), 'Moral Worth', *The Journal of Philosophy*, vol.99, no.5, pp223-245

Aylsworth, T. (2020), 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising', *Journal of Business Ethics*, vol.175, no.4, pp.689-699

Baghramian, M. (2004), *Relativism*, London, Routlege

Barton, A. and Grüne-Yanoff, T. (2015), 'From Libertarian paternalism to Nudging – and Beyond', *Review of Philosophy and Psychology*, vol.6, no.3, pp.342-359

Baxley, A.M. (2003), 'Does Kantian Virtue Amount to more than Continence?', *The Review of Metaphysics*, vol.56, no.3, pp.559-586

Baxley, A.M. (2010), *Kant's Theory of Virtue*, New York, Cambridge University Press

BBC News (2019), 'General Election 2019: Source of UK-US trade document leak must be found – PM', *BBC*, Available at https://www.bbc.co.uk/news/uk-50699168 [webpage] (accessed 12/12/2019)

Bennett, J. (1974), 'The Conscience of Huckleberry Finn', *Philosophy*, vol.49, no.188, pp.123-134

Berlin, I. (2002 [1957]), 'Two Concepts of Liberty' in Hardy, H. [ed.] (2002), *Liberty*, Oxford, Oxford University Press, pp.166-217

Berman, M.N. (2013), 'Rehabilitating Retributivism', *Law and Philosophy*, vol.32, no.1, pp.83-108

Bex-Priestly, G. (2018), 'Error and the Limits of Quasi-realism', *Ethical Theory and Moral Practice*, vol.21, no.5, pp.1051-1063

Bhatt, M.A. (2012), 'Evaluations and Associations: A Neural-network Model of Advertising and Consumer Choice', *Journal of Economic Behaviour and Organisation*, vol.82, no.1, pp.236-255

Biegler, P. and Vargas, P. (2016), 'Feeling is Believing: Evaluative Conditioning and the Ethics of Pharmaceutical Advertising', *Journal of Bioethical Inquiry*, vol.13, no.2, pp.271-279

Bollinger, L. (2002), 'Say it, Jim: the Morality of Connection in Adventures of Huckleberry Finn', *College Literature*, vol.29, no.1, pp.32-52

Bowen, J. (2011), 'The Life of Dickens I: before Ellen ternan' in Ledger, S. and Furneaux, H. (2011) *Charles Dickens in Context*, Cambridge, Cambridge University Press, pp.3-10

Brady, W.J., Crocket, M.J., and Van Bavel, J.J. (2020), 'The MAD Model of Moral Contagion: The Role of Motivation, Attention and Design in the spread of Moralised Content Online', *Perspectives on Psychological Science*, vol.15, no.4, pp.978-1010

Brandt, R. (1950), 'Stevenson's Defence of the Emotive Theory', *The Philosophical Review*, vol.59, no.4, pp.535-540

Brandt, R. (1950), 'The Emotive Theory of Ethics', *The Philosophical Review*, vol.59, no.3, pp.305-318

Brandt, R. (1979), *A Theory of the Good and the Right*, New York, Oxford University Press

Brown, H.C. (1929), 'Advertising and Propaganda: A Study in the Ethics of Social Control', *International Journal of Ethics*, vol.40, no.1, pp.39-55

Brown, W. (1997), 'Triumph of the Will', *History Today*, vol.47, no.1, pp.24-28

Bublitz, C. (2016), 'Moral Enhancement and Mental Freedom', *Journal of Applied Philosophy*, vol.33, no.1, pp.88-104

Buchwitz, L.A. (2018), 'A Model of Periodization of Radio and Internet Advertising History', *Journal of Historical Marketing Research*, vol.10, no.2, pp.130-150

Calhoun, L. (2001), 'At What Price Repentance? Reflections on Kubrick's *A Clockwork Orange*', *Journal of Thought*, vol.36, no.1, pp.17-34

Carbonell, V. (2013), 'De Dicto Desires and Morality as Fetish', *Philosophical Studies*, vol.163, no.2, pp.459-477

Carston, R. and Uchida, S. (1998), *Relevance Theory: Applications and Implications*, Amsterdam, John Benjamins Publishing Company

Caruso, F., Giuffrida, G. and Zarba, C. (2015), 'Heuristic Bayesian Targeting of Banner Advertising', *Optimisation and Engineering*, vol.6, no.1, pp.247-257

Cavendish, R. (2001), 'Publication of Uncle Tom's Cabin', *London: History Today*, vol.51, no.6, p.54

Chriss, J.J. (2014), 'Influence, Nudging and Beyond', *Society*, vol.53, no.1, pp.89-96

Christman, J. (1991), 'Autonomy and Personal History', *Canadian Journal of Philosophy*, vol.21, no.1, pp.1-24

Colingnon, S. (2018), 'Negative and Positive Liberty and the Freedom to Choose in Isaiah Berlin and Jean-Jacques Rousseau', *The Journal of Philosophical Economics*, vol.12, no.1, pp.36-64

Collins, B.M. and Stam, H.J. (2015), 'Freemans Transorbital Lobotomy as an Anomaly: A Material Culture Examination of Surgical Instruments and Operative Spaces', *History of Psychology*, vol.18, no.2, pp119-131

Cooper, R. (2014), 'On Deciding to Have a Lobotomy', *Medicine, Health Care and Philosophy*, vol.17, no.1, pp.143-154

Cowley, C. (2005), 'Changing One's Mind on Moral Matters', *Ethical theory and Moral Practice*, vol.8, no.3, pp.277-290

Crisp, R. (1987), 'Persuasive Advertising, Autonomy and the Creation of Desire', *Journal of Business Ethics*, vol.6, no.5, pp.413-418

Crossley, M.L. (2002), 'Resistance to health promotion: a preliminary comparative investigation of British and Australian students', *Health Education*, vol.102, no.6, pp.289-299

Crutchfield, P. (2016), 'The Epistemology of Moral Bioenhancement', *Bioethics*, vol.30, no.6, pp389-396

Curtin, D. and Meijer, A.J. (2006), 'Does Transparency Strengthen Legitimacy?', *Information Polity*, vol.11, no.2, pp.109-122

Cyr, T.W. and Flumer, M.T. (2018), 'Free Will, Grace and Anti-Pelagianism', *International Journal for the Philosophy of Religion*, vol.83, no.2, pp.183-199

Dai, Y. and Luqiu, L. (2020), 'Camoflaged Propaganda: A Survey Experiment on Political Native Advertising, *Research and Politics*, vol.7, no.3, pp.1-10

Davis, S.T. (1991), 'Pascal on Self-caused Belief', *Religious Studies*, vol.27, no.1, pp.27-37; p.28

De Brigard, F. (2010), 'If You Like it, Does it Matter if it's Real?', *Philosophical Psychology*, vol.23, no.1, pp.43-57

De Dreu, C.K.W., Greer, L.L., Van Kleef, G., Shalvi, S. and Handgraaf, M.I.J (2011), 'Oxytocin Promotes Human Ethnocentrism', *Proceedings of the National Accademy of Sciences of the United States of America*, vol.108, no.4, pp.1262-1266

De La Escosura, L.P. (2015), 'Human Development as Positive Freedom: Latin America in Historical Perspective', *Journal of Human Development and Capabilities*, vol.16, no.3, pp.342-373

DeCook, J.R. (2018), 'Memes and Symbolic Violence: #proudboys and the use of Memes for the Propaganda and the Construction of Collective Identity', *Learning, Media and Technology*, vol.43, no.4, pp.485-504; p.490

Dimova-Cookson, M. (2013), 'Defending Isaiah Berlin's Distinctions between Positive and Negative Freedoms' in Baum, B. and Nichols, R. [eds.] (2013), *Isaiah Berlin and the Politics of Freedom: 'Two Concepts of Liberty' 50 Years Later*, New York, Routledge, pp.73-126

Dorsey, D. (2016), 'Moral Distinctiveness and Moral Inquiry', *Ethics*, vol.126, no.3, pp.747-773

Douglas, T. (2013), 'Moral Enhancement via Direct Emotion Modulation: A Reply to John Harris', *Bioethics*, vol.27, no.3, pp.160-168

Douglas, T. (2014), 'Criminal Rehabilitation through Medical Intervention: Moral Liability and the Right to Bodily Integrity', *The Journal of Ethics*, vol.18, no.2, pp.101-122

Douglas, T. (2014), 'Moral Bioenhancement, Freedom and Reasoning', Journal of Medical Ethics, vol.40, no.6, pp.359-360

Drumwright, M.E. and Murphy, P.E. (2009), 'The Current State of Advertising Ethics: Industry and Academic Perspectives', *Journal of Advertising*, vol.38, no.1, pp.83-108

El-Enany, N. (2016), 'Aylan Kurdi: The Human Refugee', *Law Critique*, vol.27, no.1, pp.13-15

Elster, J. (1977), 'Ulysses and the Sirens: A Theory of Imperfect Rationality', *Social Science Information*, vol.16, no.5, pp.469-526

Engelen, B. and Nys, T. (2020), 'Nudging and Autonomy', *Review of Philosophy and Psychology*, vol.11, no.1, pp.137-156

Ferrara, F. (2017), 'Contagion Dynamics of Extremist Propaganda on Social Networks', *Information Sciences*, vol.418-419, no.1, pp.1-12

Fetterman, A.K., Robinson, M.D. and Meier, B.P. (2012), 'Anger as "Seeing Red': Evidence for a Perceptual Association', *Cognition and Emotion*, vol.26, no.8, pp.1445-1458

Foot, P. (2002), *Moral Dilemmas and Other Topics in Philosophy*, Oxford, Oxford University Press

Frankfurt, H. (1971), 'Freedom of the Will and the Concept of a Person', *The Journal of Philosophy*, vol.68, no.1, pp.5-20

Fransen, M.L., Smit, E.G. and Peeter, W.J. (2015), 'Strategies and Motives for Resistance to Persuasion: an Integrative Framework', *Frontiers in Psychology*, vol.6, article 1201

Geberich, W.W., Ballarini, R., Hintsala, E.D., Mishra, M., Molinari, J. and Szlufarska, I. (2015), 'Toward Demystifying the Mohs Hardness Scale', *Journal of the American Ceramic Society*, vol.98, no.9, pp.2681-2688

Gibbard, A. (1990), *Wise Choices, Apt Feelings*, New York, Oxford University Press

Grau, S.L. and Zotos, Y.C. (2016), 'Gender Stereotypes in Advertising: A Review of Current Research', *International Journal of Advertising*, vol.35, no.5, pp.761-770

Hagman, W., Andersson, D., Västfjäll, D. and Tinghög, G. (2015), 'Public Views on Policies Involving Nudges', *Review of Philosophy and Psychology*, vol.6, no.3, pp.439-453

Hagood, T.C. (2012), '"Oh what a slanderous book": Reading Uncle Tom's Cabin in the Antebellum South', *Southern Quarterly*, vol.49, no.4, pp.71-93

Hands, D.W. (2013), 'Foundations of Contemporary Revealed Preference Theory', *Erkenntnis*, vol.78, no.5, pp.1081-1108

Harris, J. (2013), ''Ethics is for the Bad Guys!' Putting the 'Moral' into Moral Enhancement', *Bioethics*, vol.27, no.3, pp.169-173

Harris, J. (2014), 'Taking Liberties with Free Fall', *Journal of Medical Ethics*, vol.60, no.6, pp.371-374

Harris, J. (2016), *How to be Good: the Possibility of Moral Enhancement*, Oxford. Oxford University Press

Hewitt, G. (2017), 'Trump and truth', *BBC*, Available at https://www.bbc.co.uk/news/world-us-canada-38731191 [webpage] (accessed 10/02/2020)

Hinds, J., Williams, E.J. and Joinson, A.N. (2020), '"It Wouldn't Happen to Me": Privacy Concerns and Perspectives following the Cambridge Analytica Scandal', *International Journal of Human-Computer Studies*, vol.19, no.8

Holmes, M. (2016), 'The 'Brainwashing' Dilemma', *History Workshop Journal*, vol.81, no.1, pp.285-293

Holton, R. (2003), 'How is Weakness of the Will Possible?' in Stroud, S. and Tappolet, C. [eds.] (2003), *Weakness of the Will and Practical Irrationality*, Oxford, Oxford University Press; pp.39-67

Hookway, C. (1998),'Doubt: Affective States and the Regulation of Inquiry', *Canadian journal of Inquiry*, vol.28, no.1, pp.203-225

Huemer, M. (2005), *Ethical Intuitionism*, New York, Palgrave Macmillan

Ibsen, H. (1882[2018]) trans. Sharp, R.F. (2018), *An Enemy of the People*, Urbana (Illinois), Project Gutenburg, available at https://www.gutenberg.org/files/2446/2446-h/2446-h.htm [accessed 14/1/2022]

Introvigne, M. (2014), 'Advocacy, Brainwashing Theories and New Religious Movements', *Religion*, vol.44, no.2, pp.303-319

Jaggar, A.M. and Tobin, T.W. (2013), 'Situating Moral Justification: Rethinking the Mission of Moral Epistemology', *Metaphilosophy*, vol.44, no.4, pp.383-408

Jensen, D. (2012), 'Kant and a Problem of Motivation', *The Journal of Value Inquiry*, vol.46, no.1, pp.83-86

Jollimore, T. (2020), 'Impartiality' in Zalta, E.N. (2021) (ed.) *Stamford Encyclopaedia of Philosophy*, Fall 2021 Edition

Kahneman, D. (2011), *Thinking Fast and Slow*, Farrar, Straus and Giroux, New York

Kant, I. (1794[1998]) trans. Wood, A and di Giovanni, G. (1998), 'Religion Within the Boundaries of Mere Reason' in Wood, A. and di Giovanni, G. (1998) [eds.], *Religion Within the Boundaries of Mere Reason and Other Writings*, Cambridge, Cambridge University Press

Kant, I. (1797[1996]) trans. Gregor, M. (1996), *The Metaphysics of Morals*, Cambridge, Cambridge University Press

Kant, I. (1785[1998]) trans. Gregor, M. (1998), *Groundwork of the Metaphysics of Morals*, Cambridge, Cambridge University Press

Kant, I. (1788[1997]) trans. Gregor, M. (1997), *Critique of Practical Reason*, Cambridge, Cambridge University Press

Kant, I., Gregor, M. and Timmerman, J. (2011), *Immanuel Kant: A Groundwork of the Metaphysics of Morals: A German-English Addition*, New York, Cambridge University Press

Kareklas, I., Brunel, F.F. and Coulter, R.A. (2014), 'Judgement is not Colour-Blind: the Impact of Automatic Colour Preference on Product and Advertising Preferences, *Journal of Consumer Psychology*, vol.24, no.1, pp.87-95

Katopol, P. (2018), 'The Halo Effect and Bounder Rationality – Limits on Decision Making', *Library Leadership and Management*, vol.32, is.3, pp.1-5

Kaufman, A.S. (1962), 'Professor Berlin on 'Negative Freedom'', *Mind*, vol.71, no.2, pp.241-243

Kelters, B. (2016), 'Gibbard's Indirect Pragmatism: Two Problems', MA Thesis, Sheffield, University of Sheffield; pp.21-22

Kennedy, G.A. (1994), *A New History of Classical Rhetoric*, Princeton, Princeton University Press

Kennedy, J.F. (1963), 'Remarks at Amherst College', Event in Honour of Robert Frost, Amherst, Massachusetts, October 26 1963

Kennett, J. (2001), *Agency and Responsibility*, New York, Oxford University Press

Kim, J., Giroux, M., Gonzalez-Jimenez, H., Jang, S., Kim, S., Park, J., Kim, J.E., Lee, J.C. and Choi, Y.K. (2020), 'Nudging to reduce the Percieved Threat of Coronavirus and Stockpiling Intention', *Journal of Advertising*, vol.49, no.5, pp.633-647

Kirkpatrick, D. (2016), 'Study: 71% of Consumers Prefer Personalised Ads', *Marketing Dive*, available at: https://www.marketingdive.com/news/study-71-of-consumers-prefer-personalized-ads/418831/ [webpage] [accessed 24/1/2021]

Klein, C.F. (2000), 'The Devil and the Skirt: An Iconographic Inquiry into the pre-Hispanic Nature of the Tzitzimime', *Ancient Mesoamerica*, vol.11, no.1, pp.1-26

Kleist, C. (2009), 'Huck Finn the Inverse Akratic: Empathy and Justice', *Ethical Theory and Moral Practice*, vol.12, no.3, pp.257-266; pp.264-265

Koenig, X. and Hilber, K. (2015), 'The Anti-Addiction Drug Ibogaine and the Heart: A Delicate Relation', *Molecules*, vol.20, no.2, pp.2208-2228

Korsgaard, C. (1996), *Creating the Kingdom of Ends*, New York, Cambridge University Press

Korsgaard, C. (1996), *Sources of Normativity*, Massachusetts, Cambridge University Press

Langsam, H. (1997), 'How to Combat Nihilism: Reflections on Nietzsche's Critique of Morality', *History of Philosophy Quarterly*, vol.14, no.2, pp.235-253

Langton, D. (2011), *Visual Marketing*, Wiley, New Jersey

Lenman, J. (2014), 'Gibbardian Fallibility: Moral Fallibility and Moral Smugness', *Journal of Value Inquiry*, vol.48, no.2, pp.235-245

Levy, N., Kahane, G., Cohen, P., Hewstone, M., and Savulescu, J. (2014), 'Are You Morally Enhanced?: The Moral Effects of Widely Used Pharmaceuticals', *Philosophy, Psychiatry and Psychology*, vol.21, no.2, pp.111-125

Luchs, F.A. (1950), 'To Religious Experience by Pathways of Art', *Religious Education*, vol.45, no.6, pp.349-352

Lutz, M. (2014), 'The 'Now What' Problem for Error Theory', *Philosophical Studies*, vol.171, no.2, pp.351-371

Mackay, A. (2005), *The Practice of Advertising [5th edition]*, Oxford, Elsevier Butterworth-Heinemann

Macmillan, M. and Lena, M.L. (2010), 'Rehabilitating Phineas Gage', *Neuropsychological Rehabilitation*, vol.20, no.5, pp.641-658

McKeon, A., Frye, M.A. and Delanty, N. (2008), 'The Alcohol Withdrawal Syndrome', *Journal of Neurology, Neurosurgery and Psychiatry*, vol.79, no.8, pp.854-862

Michelmore, M. (2012), *Tax and Spend*, Philadelphia, University of Philadelphia Press

Milevski, V. (2017), 'Weakness of Will and Motivational Internalism', *Philosophical Psychology*, vol.30, no.1, pp.44-57

Mills, C. (1995), 'Politics and Manipulation', *Social Theory and Practice*, vol.21, no.1, pp.97-112

Möller, K. (2012), *The Global Model of Constitutional Rights*, Oxford, Oxford University Press

Montminy, M. (2018), 'Culpability and Irresponsibility', *Criminal Law and Philosophy*, no.12, pp.167-181

Moriarty, M. (2020), *Pascal: Reasoning and Belief*, Oxford, Oxford University Press

Morriston, W. (2009), 'What if God Commanded Something Terrible? A Problem for Divine Command Metaethics', *Religious Studies*, vol.45, is.3, pp249-267

Mukhopadhyay, A. and Johar, G.V. (2007). 'Tempted or Not? The Effect of Recent Purchase History on Responses to Affective Advertising', *Journal of Consumer Research*, vol.33, no.4, pp.445-453

Nelson, M.R. (2008), 'The Hidden Persuaders: Then and Now', *Journal of Advertising*, vol.37, no.1, pp113-126

Nix, A. (2016), 'The Power of Big Data and Psychographics' [presentation], 2016 Concordia Annual Summit, New York, available at https://www.youtube.com/watch?v=n8Dd5aVXLCc [video] [accessed 2/8/2018]

Noggle, R. (1996), 'Manipulative Actions: A Conceptual and Moral Analysis', *American Philosophical Quarterly*, vol.33, no.1, pp.43-55

Noggle, R. (2018), 'Manipulation, Salience and Nudges', *Bioethics*, vol.32, no.3, pp.164-170

Noggle, R. (2020), 'Manipulation: A Unified Account', *American Philosophical Quarterly*, vol.57, no.3, pp.251-252

Norcross, A. (2008), 'Off Her Trolley? Frances Kamm and the Metaphysics of Morality', *Utilitas*, vol.20, no.1, pp.65-80

Nozick, R. (1972), *Anarchy, State and Utopia*, New York, Basic Books

Olson, J. (2002), 'Are Desires De Dicto Fetishistic?', *Inquiry*, vol.45, no.1, pp.89-96

Parfait, C. (2016), *The Publishing History of Uncle Tom's Cabin*, London, Routledge

Persson, I. and Savulescu, J. (2012), *Unfit for the Future*, Oxford, Oxford University Press

Persson, I. and Savulescu, J. (2015), 'Summary of Unfit for the Future', *Journal of Medical Ethics*, vol.41, no.4, pp.338-339

Pettit, P. (1999), *Republicanism: A Theory of Freedom and Government*, Oxford, Oxford University Press

Pettit, P. (2001), *A Theory of* Freedom, Cambridge, Blackwell

Pettit, P. (2006), 'Freedom in the Market', *Politics, Philosophy and Economics*, vol.5, no.2, pp.131-149

Pianalto, M. (2011), 'Moral Conviction', *Journal of Applied Philosophy*, vol.28, no.4, pp.381-395

Prozorov, S. (2019), 'Why is there Truth? Foucault in an Age of Post-truth Politics', *Constellations*, vol.26, no.1, pp.18-30

Rabosi, E. (2003), 'Some notes on Neurath's ship and Quine's sailors', *Principia*, vol.7, no.1, pp.171-184

Railton, P. (1986), 'Facts and Values', *Philosophical Topics*, vol.14, no.2, pp.5-31

Raja, W., Anand, S. and Allan, D. (2020), 'How Ad Music Attitude-Based Customer Segmentation can help Advertisers', *Journal of International Consumer Marketing*, vol.32, no.5, pp.383-399

Rakić, V. (2017), 'Compulsory Administration of Oxytocin does not Result in Genuine Moral Enhancement', *Medicine, Health Care and Philosophy*, vol.20, no.3, pp.291-297

Ramirez, E. (2016), 'Neurosurgery for Psychopaths? The Problems of Empathy and Neurodiversity', *AJOB Neuroscience*, vol.7, no.3, pp.166-168

Richardson, J.T. and Introvigne, M. (2001), '"Brainwashing" Theories in European Administrative and Parliamentary Reports on "Cults" and "Sects"', *Journal for the Scientific Study of Religion*, vol.40, no.2, pp.143-168

Ridge, M. (2003), 'Non-Cognitivist Pragmatics and Stevenson's 'Do so as well!', *Canadian Journal of Philosophy*, vol.33, no.4, pp.563-574

Rogers, K.A. (2004), 'Augustine's Compatibalism', *Religious Studies*, vol.40, no.4, pp.415-435

Rorty, A.O. (1980), 'Where does the Akratic Break take Place?', *Australasian Journal of Philosophy*, vol.58, no.4; pp.333-346

Rosati, C. (1996), 'Internalism and the Good for a Person', *Ethics*, vol.106, no.2, pp.293-326

Rosati, C. (2000), 'Brandt's Notion of Therapeutic Agency', *Ethics*, vol.110, no.4, pp.780-811

Rowland, R. (2017), 'The Significance of Fundamental Moral Disagreement', *Noûs*, vol.51, no.4, pp.802-831

Savulescu, J and Persson, I. (2012), 'Moral Enhancement, freedom and the God Machine', *The Monist*, vol.95, no.3, pp.399-421

Sabini, J. and Silver, M. (2006), 'Lack of Character? Situationalism Critiqued', *Ethics*, vol.115, no.3, pp.535-562

Samuel, L.R. (2013), *Freud on Madison Avenue: Motivation Research and Subliminal Advertising in America*, Philadelphia, University of Pennsylvania Press

Sasaki, T. (2011), 'Major Twentieth-century Critical Responses' in Ledger, S. and Furneaux, H. (2011) *Charles Dickens in Context*, Cambridge, Cambridge University Press, pp.51-58

Scanlon, T.M. (2011), 'Why not Base Free Speech on Autonomy or Democracy?', *Virginia Law Review*, vol.97, no.3, pp.541-548

Scanlon, T.M. (2014), *Being Realistic about Reasons*, Oxford, Oxford University Press

Schmidt, S. and Eisend, M. (2015), 'Advertising Repetition: A Meta-Analysis on Effective Frequency in Advertising', *Journal of Advertising*, vol.44, no.4, pp.415-428

Schroeder, M. (2007), *Slaves of the Passions*, Oxford, Oxford University Press

Schumski, I. (2017), 'The Problem of Relevant Descriptions and the Scope of Moral principles', *European Journal of Philosophy*, vol.25, no.4, pp.1588-1613

Sellars, R.W. (1957), 'Guided Causality, Using, and "Free Will"', *The Journal of Philosophy*, vol.54, no.16, pp.485-493

Severová, L., Kopecká, L., Svoboda, R. and Brčák, J. (2011), 'Oligopoly Competition in the Market with Food Products', *Agricultural Economics (Praha)*, vol.57, no.12, pp.580-588

Sgarbi, M. (2012), *Kant on Spontaneity*, London, Bloomsbury

Shapiro, T. (2009), 'The Nature of Inclination', *Ethics*, vol.119, no.2, pp.229-256

Sikka, S. (2012), 'Moral Relativism and the Concept of Culture', *Theoria*, vol.15, no.1, pp50-59

Skinner, B.F. (1971), *Beyond Freedom and Dignity*, London, Penguin Books

Smith, L.G.E., McGarty, C., and Thomas, E.F. (2018), 'After Aylan Kurdi: How Tweeting About Death, Threat and Harm Predict Increased Expressions of Solidarity With Refugees Over Time', *Psychological Science*, vol.29, no.4, pp.623-634

Smith, M. (1994), *The Moral Problem*, Malden, Blackwell

Sobel, D. (1994), 'Full Information Accounts of Well-Being', *Ethics*, vol.104, no.4, pp.784-810

Sobel, D. (2001), 'Subjective Accounts of Reasons for Action', *Ethics*, vol.111, no.3, pp.461-492

Specker, J., Focquaert, F., Raus, K., Sterckx, S. and Schermer, M. (2014), 'The Ethical Desirability of Moral Bioenhancement: A Review of Reasons', *BMC Medical Ethics*, vol.15, no.1; p.67

Specker, J., Schermer, M.H.N., and Reiner, P.B. (2017), 'Public Attitudes Towards Moral Enhancement. Evidence that Means Matter Morally', *Neuroethics*, vol.10, no.3, pp.405-417

Steinbock, B. (1981), 'Moral Reasons and Relativism', *The Journal of Value Inquiry*, vol.15, no.2, pp.157-168; pp.158

Stevenson, C.L. (1944), *Ethics and Language*, London, Yale University Press

Stevenson, C.L. (1950), 'Brandt's Questions about Emotive Ethics', *The Philosophical Review*, vol.59, no.4, pp.528-534

Stone, J. (2013), ''Unlucky' Gettier Cases', *Pacific Philosophical Quarterly*, vol.94, no.3, pp421-430

Stowe, H.B. (1852[2016]), *Uncle Tom's Cabin*, Salt Lake City, Project Gutenburg Literary Archive Foundation, available at http://www.gutenberg.org/files/203/203-h/203-h.htm [accessed 17/11/2017]

Sunstein, C. (2014), 'Nudging: A Very Short Guide', *Journal of Consumer Policy*, vol.37, no.4, pp.583-588

Sunstein, C. (2015), 'The Ethics of Nudging', *Yale Journal on Regulation*, vol.32, no.2, pp.413-450

Sunstein, C. (2016), *The Ethics of Influence*, New York, Cambridge University Press

Tait, S. (2011), 'Bearing Witness, Journalism and Moral Responsibility', *Media, Culture & Society*, vol.33. no.8, pp.1220-1235

Tiberius, V. (2012), 'Open-Mindedness and Normative Contingency' in Shafer-Landau, R. (2012), *Oxford Studies in Metaethics Volume 7*, Oxford, Oxford University Press

Tiberius, V. (2018), *Well-being as Value Fulfilment: How we can Help Each Other to Live Well*, Oxford, Oxford University Press

Tiley, J.J. (2004), 'Justifying Reasons, Motivating Reasons and Agent Relativism in Ethics', *Philosophical Studies*, vol.118, no.3, pp.373-399

Timmermann, J. (2009), *Kant's Groundwork of the Metaphysics of Morals*, New York, Cambridge University Press

Tocker, G. (2011), 'The Risk of the Nanny State: Lifestyle Advice in Public Health Campaigns', *Arbeiten aus Anglistik und Amerikanstik*, vol.36, no.1, pp.3-16

Tone, A. and Koziol, M. (2018), '(F)ailing women in psychiatry: lessons in a painful past', *Canadian Medical Association Journal*, vol.190. is.20, pp.624-625

Velleman, J. (1988), 'Brandt's Definition of 'Good'', *The Philosophical Review*, vol.97, no.3, pp.353-371

Villarán, A. (2017), 'Irrational Advertising and Moral Autonomy', *Journal of Business Ethics*, vol.144, no.3, pp.479-490

Wachowski, L. and Wachowski, A. (1999), *The Matrix*, Film, Los Angeles, Warner Brothers

Wallis, L. (2013), 'Scared Smokeless: Graphic Antismoking Ads Increase Quitting Attempts', *The American Journal of Nursing*, vol.13, no.2, p.16

Watson, C.S. (1976), 'Simm's Review of Uncle tom's Cabin', *American literature*, vol.48, no.3, pp.365-368

Wheatley, T. (2009), 'Everyday Confabulation' in Hirstein, W. [eds.] (2009) *Confabulation*, Oxford, Oxford University Press; pp.203-222

Williams, B. (1985), *Ethics and the Limits of Philosophy*, London, HarperCollins

Won, S. and Westland, S. (2017), 'Colour Meaning and Context', *Colour Research & Application*, vol.42, no.4, pp.450-459

Yaffe, G. (2003), 'Indoctrination, Coercion and Freedom of the Will', *Philosophy and Phenomenological Research*, vol.67, no. 2, pp.335-356

Zaki, J., Weber, J., Bolger, N., Ochsner, K., and Posner, M.I. (2009), 'The Neural Bases of Empathic Accuracy', *Proceedings of the National Academy of Sciences of the United States of America*, vol.106, no.27, pp.11382-11387

Zimmerman, A. (2010), *Moral Epistemology*, New York, Routledge