3

Controlling Influences

Recall that the standard account of autonomy states that autonomy requires the absence of both internal and external controlling forces that determine the agent's decision. As I suggested in the introductory chapter, the claim that autonomy requires the absence of *internal* controlling forces is too strong if it is understood to foreclose the possibility of compatibilist approaches to autonomy. Accordingly, I have argued that we should replace the condition concerning the absence of internal control with a rationalist authenticity condition. In this chapter, I now want to consider the implications of this account for the *external* forms of controlling influence to which the standard account appeals: manipulation, deception, and coercion.[1]

Although it is all but universally agreed that manipulation, deception, and coercion can undermine an agent's decisional autonomy, it is less clear how we ought to distinguish these forms of influence from those that are compatible with decisional autonomy. After all, our decisions are continually and (as relational theorists of autonomy stress) unavoidably influenced in a number of ways that are not aptly construed as undermining decisional autonomy.

To illustrate, suppose I tell you that you ought to buy a novel, telling you that it incorporates beautiful prose and cutting social commentary. *Ceteris paribus*, it does not seem that this way of attempting to influence your decision is aptly construed as undermining your autonomy with regards to your decision about which book to buy. Even at a pre-theoretic stage, this clearly stands in stark contrast to a case in which I threaten to significantly harm you if you do not read the book. Consider also a case in which I maliciously deceive you into reading a book that I know you will dislike, or, even more fantastically, hypnotizing you into doing so.

Accordingly, the challenge that we face in understanding controlling influence is to explain how we are to distinguish those forms of influence that serve to undermine an individual's decisional autonomy from those that do not. Although the standard account stipulates a condition regarding the absence of controlling interference, the

[1] 'Undue influence' is arguably a more natural term for what I am referring to as controlling influence. However, I have avoided this term due to the fact that, in some circles it is taken to have a rather more specific meaning than I intend. For example, the Belmont Report defines undue influence as follows: 'Undue influence . . . occurs through an offer of an excessive, unwarranted, inappropriate or improper reward or other overture in order to obtain compliance.' See Largent et al., 'Misconceptions about Coercion and Undue Influence' for discussion. I shall discuss the implications of incentives for autonomy in the next chapter.

Autonomy, Rationality, and Contemporary Bioethics. Jonathan Pugh, Oxford University Press (2020). © Jonathan Pugh.
DOI: 10.1093/oso/9780198858584.001.0001

account itself provides us with few clues about how to draw these distinctions; it simply stipulates by fiat that coercion, psychological manipulation, and deception are examples of interventions that undermine autonomy. It relies on the intuitive plausibility of these judgements in order to justify partially defining autonomous decisions as those that are made in the absence of these influences. This seems somewhat theoretically unsatisfactory.

The rationalist approach that I developed in the previous chapter can offer a deeper account of the relationship between autonomy and these different forms of influence. On this approach, an individual's decisional autonomy can be undermined if either the cognitive or reflective elements of decisional autonomy are disrupted as follows:

(i) The individual is led to either (a) sustain theoretically irrational beliefs or (b) fail to hold decisionally necessary true beliefs.

(ii) The individual is led to sustain a motivating desire in a manner that bypasses the cognitive element of autonomy, such that they either (a) do not endorse the desire, or (b) they endorse it with a preference that fails to cohere with other cohering elements of their character system

In accordance with this framework, we may categorize deception as amounting to (i)[b] but not necessarily (i)[a]. It can be theoretically rational to believe others who are lying to us. However, informational manipulation more broadly may involve either (i)[a] or [b]. I shall consider these forms of controlling influence in sections 4 and 5.

In contrast, *psychological* manipulation involves the manipulation of motivational states rather than beliefs, and may be said to amount to forms of influence that are involved in either (ii)[a] or [b]. One might argue that psychological manipulation could also operate at a more global level, such that an agent is brainwashed into holding an entirely new character system as follows:

(ii)[c] An individual is led to radically change the overall coherent nexus of preferences and acceptances by which she endorses her motivating desires.

Whilst the theory of autonomy that I outlined in the previous chapter can explain why (ii)[a] and [b] undermine autonomy, it is not clear that it can explain why global manipulation of the sort identified in (ii)[c] would. As I shall explain, this has led some philosophers to argue that an adequate theory of autonomy should certain incorporate historical conditions.

I shall reject this view in section 3. Prior to doing so though, I shall in the first two sections begin by outlining more mundane forms of psychological manipulation that may more plausibly be employed in biomedical contexts, and explaining how they differ from forms of rational persuasion that are compatible with decisional autonomy. I shall delay consideration of coercion until the next chapter, and explain why it does not fit naturally within the framework I have just outlined.

To conclude these introductory remarks, it is important to be clear that I am only interested here in the implications of these forms of influence for the *autonomy* of the targeted individuals. I do not mean to deny that these influences can have other important moral implications that can and should factor into an all things considered moral analysis of them.[2] For instance, there has been a great deal of recent debate regarding the extent to which instances of psychological manipulation might violate a putative right to mental integrity, even if they do not constitute threats to autonomy per se.[3] Although my autonomy-based analysis of manipulation is not entirely unrelated to these other questions, we should not assume that the conclusions I draw here translate straightforwardly to how we should understand other important moral properties of manipulation and deception.

Finally, the terminology of 'controlling influence' is perhaps somewhat unfortunate, since it seems to imply that these forms of influence must be exerted intentionally, in a manner that connotes that the one who influences is actively *controlling* the target of their influence. However, as I shall explain in this chapter, non-intentional and indeed non-agential forces can cause the phenomena indicated in (i) and (ii) above. Despite this unfortunate implication, I shall retain this terminology in the interests of consistency with the literature.

1. Rational Persuasion

On the standard account, rational persuasion is understood to be compatible with autonomy on the basis that it enhances understanding; persuasion is thus broadly similar to simply informing.[4] However, on the approach that I outlined in the previous chapter, we can understand persuasion in a broader sense to involve attempting to change an agent's beliefs by drawing their attention to reasons. Rational persuasion on this approach thus involves attempting to change an individual's desires indirectly, by actively engaging with the cognitive element of their practical decision-making process. This can include interacting with both the target's descriptive beliefs about the world (as the standard account implies), but also their evaluative beliefs about the good. Desires that are formed following rational persuasion are thus likely to accord with the cohering elements of the agent's character system in a manner that betokens autonomy, because they will have been explicitly developed in light of the individual's acceptances. When successful, persuasion must appeal to the values we hold, or convince us to change our values in response to reasons.

Typically, rational persuasion will involve highlighting descriptive facts about another's options; for example, you might persuade a friend not to cross a bridge by drawing their attention to the large hole in the middle of it. We may call this factual persuasion. In factual persuasion, the persuader presumes that both they and the subject of their persuasion share an understanding of the target's preferred outcomes. One explanation for why factual persuasion can fail is that the persuader has made an incorrect assumption about the target's preferred outcome.

[2] For a comprehensive discussion, see Blumenthal-Barby, 'A Framework for Assessing the Moral Status of "Manipulation"'.

[3] Bublitz and Merkel, 'Crimes Against Minds'; Douglas, 'Neural and Environmental Modulation of Motivation'.

[4] Faden and Beauchamp, *A History and Theory of Informed Consent*, 354–68; Beauchamp and Childress, *Principles of Biomedical Ethics*, 137–9.

Alternatively, the two parties may agree on all the relevant descriptive facts of the matter but disagree about the outcome that ought to be pursued in a particular context. In such cases, simply bringing further relevant descriptive facts to the target's attention is unlikely to be successful. Rather, if persuasion is to succeed in such a context, it must involve what I shall call 'evaluative persuasion'; that is, the persuader must bring reason-giving facts about other outcomes to the target's attention, in an attempt to change their assessment of the relative strength of their reasons to pursue different courses of action.

Evaluative persuasion might involve advocating the value of goods that the target will forgo if they follow through on their planned course of action; for example, we might seek to persuade a suicidal person against their planned course of action by drawing their attention to various good things in their life that are worth living for. Alternatively, evaluative persuasion might involve questioning the value of the agent's preferred outcome, asking her to explain the grounding of her belief that the outcome in question is good in a reason-implying sense. For example, one might attempt to persuade a person to stop smoking by asking her to reflect on whether she actually enjoys the experience of smoking, or whether she simply reaches for her cigarettes on a habitual, non-rational basis.

In the context of medical ethics, evaluative persuasion is often viewed with suspicion, particularly by those who endorse the so-called shared decision-making model of the doctor–patient relationship.[5] On strong versions of this view, it is assumed that:

> The physician should objectively answer questions but should avoid influencing the patient to take one path or another, even if the physician has strong opinions or if the patient asks for advice.[6]

On this model, it is usually assumed that physicians would be exerting controlling influence if they were to encroach on the evaluative domain of the decision-making process. The General Medical Council (GMC) adopts a weaker version of this position, stating that:

> The doctor may recommend a particular option which they believe to be best for the patient, but they must not put pressure on the patient to accept their advice.[7]

Whilst there are different ways of understanding the shared decision-making model, we should reject versions of the model that construe the doctor–patient relationship in a manner that demands that doctors should adopt a wholly value-neutral approach in their dealings with patients.[8] First, it is highly questionable to assume that it is even possible for doctors to provide medical information in a value-neutral manner. After all, medical concepts such as health and disease are themselves value-laden, in so far as they are commonly understood to imply certain value judgements

[5] Veatch, 'Abandoning Informed Consent'.

[6] Quill and Brody, 'Physician Recommendations and Patient Autonomy', 764.

[7] General Medical Council, 'Ethical Guidance for Doctors, Part 1'.

[8] Veatch, 'Abandoning Informed Consent'. For an understanding of the doctor–patient relationship that corresponds to the rationalist approach that I am defending here, see Savulescu, 'Liberal Rationalism and Medical Decision-Making'.

(particularly in the conversational context of a treatment discussion).[9] Second, physicians must make certain evaluative judgements in deciding upon what information to disclose to patients (such as information about risks associated with treatment) and deciding what treatment options to propose to their patient. In both cases, it seems that the physician's evaluative judgements will inevitably bear upon what is (and is not) disclosed.

However, the theory of autonomy that I am outlining also gives us reasons to reject even the weaker version of this value-neutral approach adopted by the GMC. The strategies of evaluative persuasion noted above should not be understood to constitute controlling influence of the sort that undermines decisional autonomy, because evaluative, as well as factual, persuasion can facilitate autonomous decision-making. Evaluative persuasion does not involve seeking to impose values on the target; rather, it involves seeking to elicit the rational justification underlying the values that are now guiding the agent's conduct, and alerting them to other reasons that are at stake in a particular choice context. In so far as this mode of persuasion causes the subject to reflect on these reasons, it can be construed as enhancing the agent's autonomy with respect to their decisions, even if it is not successful.[10] Indeed, part of the physician's role can be to advocate the import of certain sorts or reasons, reasons that reflect the values that shape the profession of medicine.[11] Moreover, in an era in which physicians have to battle with the widespread distribution of misinformation about medicine proliferating in the online world, it may be a mistake to assume that the dispassionate provision of medical facts alone can be enough to allow them to adequately compete in this environment, and to ensure that their patients are appropriately informed about their treatment options.[12]

Of course, a decision to engage in evaluative persuasion must be sensitive to the particular vulnerabilities of specific patients. In particular, it is important to be clear that the physician is engaging in evaluative persuasion rather than simply aiming to elicit the patient's capitulation to their view. Whilst this is an important danger, we should not simply assume that the best way to avoid it is to require that doctors say nothing in the face of patient decisions that seem to be grounded by badly skewed evaluative judgements. To illustrate, a physician would be quite warranted in engaging in evaluative persuasion to persuade a patient that she really ought to receive a life-saving shot to prevent an anaphylactic shock, if her reason for refusing it is that she wants to avoid the small pain involved in the injection. A physician can go beyond merely recommending the injection in this sort of case without unduly encroaching on the patient's autonomy.

Having outlined forms of persuasion that are compatible with autonomous decision-making, I now want to turn to forms of influence that are not, starting with psychological manipulation.

[9] For a classic account of conversational implicature, see Grice, *Studies in the Way of Words*.

[10] For defence of similar views, see Savulescu, 'Liberal Rationalism and Medical Decision-Making'; Widdershoven and Abma, 'Autonomy, Dialogue, and Practical Rationality'.

[11] Brock, *Life and Death*, ch. 2, especially 69.

[12] For a sobering editorial on this point, see Ranjana Srivastava 'My Patient Swapped Chemotherapy for Essential Oils. Arguing Is a Fool's Errand'.

2. Psychological Manipulation

In rational persuasion, one attempts to alter the target's motivational states indirectly by engaging with the cognitive element of their decisional autonomy, including their beliefs about the world and the good. Psychological manipulation involves attempting to directly alter the motivational states themselves, in a manner that bypasses the cognitive element of the target's decisional autonomy.

As Anne Barnhill notes, one of the difficulties in theorizing about manipulation is that ethicists often fail to provide a working definition of what they are talking about, or they offer definitions that are either over-inclusive or under-inclusive.[13] Indeed, one over-inclusive theory that Barnhill adverts to is an account of psychological manipulation that Tom Beauchamp and Ruth Faden develop in outlining a detailed version of the standard account of autonomy in bioethics. According to Beauchamp and Faden, psychological manipulation can be defined as:

> any intentional act that successfully influences a person to belief or behavior by causing changes in mental processes other than those involved in understanding.[14]

This account of psychological manipulation is theoretically incomplete because the link between this type of manipulation and the explanation for why it undermines autonomy is left unclear. This is particularly problematic because there are ways in which the definition is *both* over- and under-inclusive. One way in which the theory may be under-inclusive is that it rules out the possibility that manipulative influence could be non-intentional. However, one reason for questioning Beauchamp and Faden's claim to the contrary is that they (and other advocates of the standard account of autonomy) readily accept that *some* non-intentional forces can amount to controlling influence that undermines decisional autonomy. This is due to the fact that advocates of the standard account accept the possibility of internal controlling forces (such as psychiatric disease) that may undermine decisional autonomy.

There are of course some pragmatic reasons for understanding manipulation to require intentional agency. First, one might want to stress the semantic point that the term 'manipulation' itself seems to connote intentional agency, in the same way that the term 'controlling' influence does. Furthermore, there may be a range of morally relevant features of intentional manipulation that are not applicable to non-intentional forms of the same sort of phenomenon. For my purposes here though, these points are somewhat moot. As I stipulated in the introduction to this chapter, I am solely interested in the effects of manipulation on the target's autonomy. Crucially, that non-intentional forces can exert forms of influence that have relevantly similar effects on the target's autonomy as intentional psychological manipulation follows straightforwardly from the account of autonomy that I have defended. If authenticity of a certain sort is required for decisional autonomy, and if authenticity as I have spelled it out can be threatened by non-intentional and non-agential forces (as well as agential forces), then we should deny the claim that autonomy-

[13] Barnhill, 'What Is Manipulation?', 51.

[14] Faden and Beauchamp, *A History and Theory of Informed Consent*, 366.

undermining psychological manipulation is necessarily intentional.[15] What matters for decisional autonomy is that we endorse our motivating desires with certain kinds of rationally endorsed preferences, namely, ones that fit in with other cohering elements of our characters. Now, it is true that we might fail to act on the basis of such a desire because another agent has intentionally induced a different motivating desire. However, it is also possible that our failure in this regard may not be due to the machinations of another intentional agent. 'Internal' forms of controlling influence of the sort I considered in the previous chapter can lead individuals to form and sustain desires in this way, and undermine autonomy for the same reason.

I shall consider some further arguments regarding the necessity of intention that have been made specifically with respect to deception in section 5. To return to psychological manipulation though, one might worry that my theory of autonomy will lead to too broad an understanding of manipulation. For instance, Barnhill argues a theory of manipulation ought to exclude drugs and brainwashing as examples of interference that undermine autonomy, on the basis that they evince *global* changes to the target's psychology. Such global changes mean that these interventions do not directly target particular elements of the target's psychological economy in the manner that manipulation arguably requires.[16] It seems plausible that psychiatric diseases could also constitute another example of what Barnhill has in mind here.

I am sympathetic to Barnhill's claim here; indeed, in the next section, I shall explain why forms of interference that evince global changes to a person's psychology do not threaten autonomy per se. However, I do not believe that these examples are problematic for the conception of manipulation that can be grounded by my rationalist approach. First, I am sceptical of the claim that drugs or psychiatric disorders must *always* involve such global changes. Indeed, some drugs have direct effects on a limited set of motivational states. Consider, for example, the use of chemical castration in the punishment of individuals who have been convicted of sexual offences. Contrary to Barnhill's analysis, this seems a paradigm case of psychological manipulation; these individuals are compelled to take a drug that has a direct effect on their libido, but which may nonetheless leave large swathes of their character systems intact. Indeed, they may lament their lack of sexual drive whilst experiencing the effects of the drug, in accordance with preferences that they have sustained from a point in time prior to the intervention. Furthermore, with respect to psychiatric diseases, we may also note that the standard account of autonomy explicitly accepts the claim that psychiatric disease can undermine decisional autonomy.

As I mentioned above, Barnhill also claims that Beauchamp and Faden's account of manipulation is over-inclusive. According to Beauchamp and Faden's approach, methods of changing beliefs and desires that do not qualify as rational persuasion will constitute psychological manipulation. Yet as Barnhill argues, it appears that intentional expressions of emotions may serve to change another's beliefs or behaviour,

[15] Mele makes a similar point with respect to his theory in Mele, *Autonomous Agents*, ch. 6, section 2 and Mele, *Free Will and Luck*, 177–90.

[16] Barnhill, 'What Is Manipulation?', 65.

but this does not entail that such expressions must be manipulative.[17] It might be argued that my approach will fall foul of a similar complaint, given the emphasis I place on the fact that psychological manipulation involves inducing desires directly by bypassing the cognitive element of autonomy. However, my approach can accommodate the thought that emotional influences can be compatible with autonomy, as I shall now explain.

One explanation for this is that emotional influences need not lead us to develop motivational states in a manner that bypasses the cognitive element of decisional autonomy. To illustrate, consider Barnhill's own example of a woman deciding to hand back some embezzled money after her father tells her that 'he didn't raise her to be a thief'.[18] I agree with Barnhill that this is plausibly an example of a non-manipulative, yet emotional form of influence (via shame) that is compatible with autonomy.

However, I deny that it involves inducing a desire in a manner that entirely bypasses the cognitive element of the agent's autonomy. It is a mistake to assume that decisions substantially grounded by affective experiences are wholly divorced from our beliefs about what is valuable. Whilst it is true that emotional experiences of fear and anger (amongst others) can involve us becoming divorced from our evaluative judgements or theoretically rational beliefs about the nature of the world, some affective attitudes and emotional experiences can instead give rise to values, ground certain kinds of reasons for choice, and even reveal the presence of certain reasons that were previously obscure to us. Indeed, an increased understanding of our reasons can sometimes also be facilitated by emotional engagement, as well as simpler forms of information disclosure. In this particular case, my claim is that the emotional influence of the woman's father enabled her to perceive a powerful set of reasons not to steal.

Of course this may not be true of all forms of emotional influence; when it is not, I suggest that emotional influence can be manipulative. However, we should also acknowledge the point that the fact that one has been initially manipulated into holding a desire (by emotional means or otherwise) does not entail that one must thereby *forever* lack autonomy with respect to it. One can come to critically reflect on the content of the manipulated desire, and to decide for oneself whether or not to *sustain* it in the light of one's preferences.

To illustrate, suppose that you were brainwashed by subliminal advertising to form the desire to give money to charity. Even if you later came to reject the causal history of this desire (having been made aware of it), it still seems plausible to claim that you could autonomously hold the desire, just because you endorse the *content* of the desire itself. As Bernard Berofsky points out, one may reasonably have objections to the causal process that led one to develop a desire, without having any qualms about the results.[19] By reflecting on a desire that one has been manipulated to develop, one can come to take ownership of the desire itself in a manner that betokens autonomy; call this 'post-factum reflection'.

[17] Ibid., 62.

[18] Ibid.

[19] Berofsky, *Liberation from Self*, 212.

Post-factum reflection on changes to our motivational states following emotional influence may often lead to their endorsement, because our emotions can plausibly serve as a source of reasons that may not be immediately accessible at the time of deliberation, but which may nonetheless ground rational behaviour.[20] More broadly though, we are of course more likely to engage in post-factum reflection on manipulated desires if they represent either a striking departure from our characteristic motivations, or if the manipulative process itself was particularly overt. The problem is that many manipulative processes, including many forms of emotional influence, subvert this sort of post-factum reflection because their effects are covert and subtle. Accordingly, they do not give us cause to reflect on, or indeed even recognize, the changes that they have evinced.

On a related point, we may note that on the approach that I am defending here is that psychological manipulation need not be covert.[21] In light of the above remarks though, my account is compatible with the claim that covert manipulation is likely to be more successful in sustaining a lack of autonomy with respect to a desire; after all, if the target of manipulation is aware of the existence of a manipulative influence they are more likely to be able to both employ mechanisms to resist its influence, and to engage in post-factum reflection that will allow them to repudiate or take ownership of the desire itself. However, it is also compatible with the claim that overt influences can still involve psychological manipulation. This seems to me a benefit of the account. To see why, reconsider the use of mandatory chemical castration in the criminal justice system.[22] Now, whatever we think of the permissibility of this sort of policy, it seems quite clear that intervention can be understood to be a form of psychological manipulation, despite the fact that it is not covert, if the target repudiates their lack of sexual desires following the intervention.

With this in mind, what might plausible forms of psychological manipulation look like in the biomedical context? Philosophers who discuss manipulation typically appeal to extraordinary cases of manipulation involving nefarious neurosurgeons and hypnotists. However, the justification for appealing to such examples is the theoretical clarity that they permit, rather than the fact that they represent common cases of manipulation. Yet, there are a number of common ways in which we might be subjected to this form of interference in the biomedical sphere. For instance, physicians may exert control in these sorts of ways through subliminal suggestion, and by appealing to irrationally grounded emotional attitudes such as guilt.[23] To illustrate, a physician could psychologically manipulate a patient who refuses a treatment by telling them that they are just 'being awkward', or by telling the patient that all of their other patients just 'do what their doctor says'. In such cases, the physician is attempting to influence the patient, not by appealing to reason-giving facts about the nature of the treatment that could give the patient reasons to change their decision, or by making an emotional appeal that serves to reveal the strength of

[20] Arpaly, 'On Acting Rationally against One's Best Judgment'.

[21] For accounts that claim that manipulation must necessarily be covert, see Goodin, *Manipulatory Politics*, 9; Ware, 'The Concept of Manipulation', 165.

[22] Forsberg and Douglas, 'Anti-Libidinal Interventions in Sex Offenders'.

[23] Faden and Beauchamp, *A History and Theory of Informed Consent*, 366–7.

the patient's reasons (such as an injunction to 'think of your family' in making a treatment decision). Instead, this is an appeal to a non-rational bias that the patient may have to conform to a 'norm' of 'the good patient' perpetuated by a medical authority.

Over the latter half of the twentieth century, psychologists and behavioural economists have also highlighted ways in which environmental cues can be strategically used to influence our practical decision-making in ways that might be deemed manipulative in various ways. Following Thaler and Sunstein, such strategies are commonly referred to as 'Nudges'. According to Thaler and Sunstein's own definition, a nudge can be constituted by:

> Any aspect of a person's choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives.[24]

The debate in the bioethical literature about whether nudges are compatible with autonomous choice has been somewhat impeded by the exceedingly broad nature of this definition of nudges. The problem with the definition is that it can be understood to incorporate both strategies that influence decision-making by facilitating the involvement of the cognitive element of our decision-making, and also those that subvert this. In doing so, it conflates two morally distinct categories of influence.[25] For instance, rational persuasion as I have described it above could qualify as a nudge on this definition; so too could the use of incentives.[26]

However, I suggest that there are some cases in which nudges are psychologically manipulative, by virtue of the fact that they entirely bypass the cognitive element of our decision-making (as well as subverting post-factum reflection). Consider for example the use of priming. In one famous example of priming, criminal justice authorities found that if they exposed prison inmates to a particular shade of pink, violent behaviour amongst those inmates dramatically reduced. It is not as if the colour prompted the inmates to engage in reflection about their reasons to engage in violent behaviour. Rather, their exposure to this environmental factor seemed to somehow serve to diminish their violent impulses without engaging with the offenders' rational processing.[27] Moreover, the covert and subtle nature of this influence makes it less likely that the targets will be aware of the changes evinced, and to critically reflect on the question of whether to endorse or reject them (assuming that the effects themselves could indeed be wilfully resisted).

Despite the breadth of the commonly invoked definition, discussions of the effects of nudging on autonomy do not always attend to the particular significance of nudges that bypass (and subvert) the rational processes alluded to above.[28] A number of commonly discussed nudge strategies, such as priming, would fall into this category

[24] Thaler and Sunstein, *Nudge*, 6.

[25] For other defences of the claim that not all nudges need pervert our decision-making processes, see Wilkinson, 'Nudging and Manipulation'; Cohen, 'Nudging and Informed Consent'.

[26] Blumenthal-Barby and Burroughs, 'Seeking Better Health Care Outcomes'.

[27] For discussion, see Pugh, 'Moral Bio-Enhancement, Freedom, Value and the Parity Principle'; Douglas, 'Neural and Environmental Modulation of Motivation'.

[28] Some authors employ a narrower definition of nudging, according to which nudges by definition take advantage of non-rational processes. For example, see Hausman and Welch, 'To Nudge or Not to Nudge'.

quite uncontroversially. However, in the case of some strategies the matter is not so straightforward. One reason for this is that some nudges involve forms of informational manipulation, rather than psychological manipulation as I have defined it here, and the distinction between informational manipulation and enhancing understanding through information disclosure can be somewhat blurred. I shall consider this form of influence in section 4.

Prior to doing so, to complete my analysis of psychological manipulation, I shall in the next section consider forms of global psychological manipulation outlined in (ii) [c] in the introduction to this chapter. Such instances of psychological manipulation have prompted some philosophers to argue that we ought to understand authenticity in a strictly historical sense. I shall defend my account from this objection, and in doing so outline how my theory can accommodate the pervasive relational influences on our values within a framework of decisional autonomy.

3. Global Manipulation and Autonomy

One might be sceptical that there is a practical need for a theory of autonomy to accommodate the prospect of global manipulation. For instance, Marilyn Friedman has argued that philosophers should refrain from engaging with bizarre counter-examples to autonomy, and that we should only refine such concepts in 'helpful practical ways'.[29] I am sympathetic to Friedman's frustration in this regard. Nonetheless, some comments on global manipulation are necessary. One reason for this is that critics of the kind of rationalist approach that I have endorsed here might contend that it is ill-equipped to accommodate the pervasive relational influences that we are all subject to as members of society. This is important because these influences arguably represent a very real way in which external forces can substantially mould and shape our character systems as a whole. As I shall explain, this sort of observation has led some relational theorists to abandon the idea that autonomy requires authenticity conditions (since authenticity would inevitably be tainted by this social influence). Second, cases of global manipulation have led other theorists to argue that authenticity should be understood in an explicitly historical sense. I shall consider each point in turn.

(i) The Pervasiveness of Relational Influence and Autonomy

It is undeniable that relational autonomy theorists have captured a number of important insights about autonomy. We are social beings, and our decisions are both guided and enabled by societal influences in a pervasive fashion.[30] Yet, these insights are compatible with a wide range of theories of decisional autonomy, including the one developed in the previous chapter.[31]

[29] Friedman, *Autonomy, Gender, Politics*, 28.

[30] For detailed discussion of the forms that relational influence can take, see the entries in the seminal Mackenzie and Stoljar, *Relational Autonomy*. See also Christman, *The Politics of Persons*, ch. 8; Oshana, 'Personal Autonomy and Society'; Nedelsky, 'Reconceiving Autonomy'.

[31] Foster suggests that relational theorists who have supposed otherwise are largely attacking a straw man. See Foster, *Choosing Life, Choosing Death*, 14–15.

In particular, most theories are compatible with the claim that certain relational conditions are *causally* necessary for autonomy.[32] This is significant, as the key claim of relational autonomy theory is sometimes understood as a claim about the *causal* conditions of autonomy. Consider for example, Anderson and Honneth's claim that:

> The key initial insight of social or relational accounts of autonomy is that full autonomy . . . is only achievable under socially supportive conditions.[33]

This is an important claim, but it is also one that can (and should) be accommodated straightforwardly by procedural theories of decisional autonomy that outline *constitutive* conditions of autonomy. In the terms of the theory that I have developed, one may maintain that decisional autonomy requires both theoretical and practical rationality whilst accepting both (i) that these capacities themselves may be socially mediated and (ii) the more fundamental point that one can only exercise these capacities in a social environment that furnishes one with the opportunities and abilities to make one's decisions in this way.[34]

The more challenging question arising from relational theory concerns whether the constitutive conditions of decisional autonomy adequately accommodate the pervasiveness of relational influences.[35] Indeed, a widespread criticism of the kinds of procedural theories of autonomy that I discussed in the previous chapter is that they focus on an unduly individualistic conception of autonomy and the self.[36]

I believe that this criticism is somewhat misplaced with regards to the rationalist theory that I have developed, since the theory is compatible with relational influence on the self and autonomy in a number of ways.[37] First, many of our practical reasons can be other-regarding. Our practical reasons are often self-interested reasons; but they can also include the reasons we might have to promote the interests of others, or reasons grounded by the value that our relationships instantiate. Accordingly, the theory is well-placed to accommodate the fact that our relationships can be of the utmost importance to us as autonomous agents.[38]

Furthermore, the claim that decisional autonomy requires practical and theoretical rationality is quite compatible with the claim that many of the values, desires, and beliefs that ground our rationality will often have been formed as a result of our relationships and social forces. Accordingly, I wholeheartedly agree with John

[32] John Christman makes a similar point. See Christman, *The Politics of Persons*, 177–82.

[33] Anderson and Honneth, 'Autonomy, Vulnerability, Recognition, and Justice', 130. For further discussions of relational causal conditions of autonomy, see also Nedelsky, 'Reconceiving Autonomy'; Mackenzie and Stoljar, 'Autonomy Reconfigured'; Oshana, 'Personal Autonomy and Society'.

[34] Nedelsky particularly stresses this point in Nedelsky, 'Reconceiving Autonomy'.

[35] Oshana similarly identifies this as the most significant challenge posed by relational influences. See Oshana, 'Personal Autonomy and Society', 97.

[36] For example, see Oshana, 'Personal Autonomy and Society'; Mackenzie and Stoljar, 'Autonomy Reconfigured'; Meyers, *Self, Society, and Personal Choice*; Anderson and Honneth, 'Autonomy, Vulnerability, Recognition, and Justice'.

[37] Christman's theory, which I criticize below, is also compatible with these claims. See Christman, *The Politics of Persons*, chs. 7 and 8. For discussions of this trend more broadly with regards to liberal conceptions of autonomy, see Friedman, *Autonomy, Gender, Politics*, 81–97.

[38] Christman, 'Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves'.

Christman when he writes that there are a number of ways that any plausible theory of autonomy must:

> . . . take into account the various ways in which humans are socially embedded, intimately related to other people, groups, institutions and histories, and that they are motivated by interests and reasons that can only be fully defined with reference to other people and things.[39]

The theory of autonomy I have outlined readily accepts these claims; the important point is whether individuals are able to reflect upon these socially mediated values in the process of cultivating their characters. It does not require that these values were developed in a social vacuum. However, the agent must take ownership of these values by ensuring that they hold their evaluative beliefs in a theoretically rational sense, and incorporate them into a coherent character system.

As I mentioned in the introductory chapter, some feminist philosophers have argued that the fact that our values have a social source is a damaging criticism of theories of autonomy that incorporate considerations of authenticity. If our 'true selves' have been uncritically forged in the crucible of a society that ensures that individuals simply internalize socially oppressive norms, then perhaps we should be sceptical of the claim that these selves are the locus of autonomous agency.[40] Instead, perhaps we ought to appeal to substantive conceptions of autonomy, or to develop non-authenticity based procedural accounts that appeal to social conditions, or competency conditions.[41]

I have already noted that relational competency conditions can be compatible with a wide range of procedural theories of decisional autonomy. However, procedural and substantive relational accounts of autonomy face a deeper conflict. The crux of the issue here is captured in the example of cosmetic surgery that I alluded to in the introduction. If one holds the view that a woman's desire for a beautifying cosmetic procedure is merely an artefact of the influence of the patriarchal society in which she lives, then one might deny that a woman can be autonomous with respect to that desire, no matter how much she personally endorses it. In this sort of example, procedural and substantive relational accounts come into irreconcilable conflict.

It would be impossible to significantly advance the debate between substantive and procedural theories on this point in the space available here. I must make do with adverting to existing work that extensively defends the procedural approach in this regard,[42] and reiterating my own concern (outlined in the introduction) that substantive accounts could legitimize paternalistic interference under the rubric of autonomy. Further, I suspect that part of the reason that this conflict appears irreconcilable is that at least some substantive theorists conceive of autonomy as something akin to a socialized ideal of what it would be live a life of independence and equal standing, rather than to live one's life in accordance with one's own

[39] Christman, *The Politics of Persons*, 165.

[40] Stoljar, 'Autonomy and the Feminist Intuition'.

[41] Mackenzie, 'Three Dimensions of Autonomy', 31. See Westlund, 'Rethinking Relational Autonomy'; Meyers, *Self, Society, and Personal Choice*; Benson, 'Autonomy and Oppressive Socialization'; Benson, 'Feminist Intuitions and the Normative Substance of Autonomy'; Stoljar, 'Autonomy and the Feminist Intuition'.

[42] Christman, *The Politics of Persons*, ch. 8; Friedman, *Autonomy, Gender, Politics*, ch. 1.

values.[43] It may be that autonomy theorists in this context are simply interested in different things that nonetheless get lumped together under the umbrella term of autonomy.

(ii) A Need for Historical Conditions?

Rather than attend further to the debate between procedural and substantive theorists, I shall instead consider whether these considerations of pervasive relational influence suggest a need to incorporate historical conditions into our understanding of what it is for a motivating desire to be authentic.

Advocates of historical theories claim that the theories of decisional autonomy that I considered in the previous chapter are *ahistorical*.[44] These theories are ahistorical in the sense that they claim that an agent's subjecting their motivating desire to a certain sort of psychological scrutiny at a particular point in time is sufficient for their being autonomous with respect to it. They are not particularly concerned about *how* the agent came to form the desire (or indeed the components of their psychology that might critically reflect on the desire). The motivation for claiming that an adequate theory of decisional autonomy should incorporate historical conditions is that agents may have been caused to have either their motivating desires, or other elements of their psychological economies in ways that appear to undermine their autonomy.

In the previous section, I explained how the account I have developed can explain why certain forms of influence constitute psychological manipulation that undermines autonomy. However, I noted that the mere fact that a desire was elicited via manipulative means does not entail that the agent must forever lack autonomy with respect to it. As I shall explain, this is a point that historical theories can also accommodate (although at some cost). However, the main point of disagreement between the theory of autonomy that I have outlined and historical approaches is that my theory cannot account for why global forms of manipulation (as outlined in (ii) [c] in the introduction to this chapter) would undermine autonomy.

One of the most widely discussed cases of global manipulation is Alfred Mele's case of Beth and Ann, in which Beth, an unproductive philosopher, is covertly brainwashed to become psychologically identical to Ann, a very productive philosopher. The brainwashers instil the same hierarchies of value that Ann has into Beth, while eradicating all of Beth's other competing values; she embraces her newfound passion for philosophy following critical reflection.[45]

Historical theorists typically use this case to object to ahistorical theories on the basis that the latter cannot account for the plausible intuition that Ann is autonomous with respect to her future philosophical behaviour in a way that Beth is not; *ex hypothesi*, both Ann and Beth (following the intervention) have identical

[43] Christman identifies this implicit conception of autonomy in some feminist discussions of autonomy in Christman, *The Politics of Persons*, 175.

[44] Mele distinguishes between internalist and externalist forms of psychological autonomy rather than ahistorical and historical forms. Mele, *Autonomous Agents*, 146–56. I have avoided the former terminology to avoid confusion with the way in which internalism and externalism have featured in debates about practical reason.

[45] Mele, *Autonomous Agents*, 145.

psychologies and thus reflectively endorse their love of hard philosophical work (a new passion in Beth's case).[46]

In contrast, one might accommodate the intuition that Beth lacks autonomy by claiming that there is either an objective or subjective historical condition of authenticity. I shall first briefly consider Mele's own approach, before turning to consider Christman's historical approach in more detail, given its recent influence in bioethical discussions.

According to Mele, a necessary condition of an agent's possessing an authentic desire is that she was not 'compelled' to have that desire, such that she is practically unable to shed it.[47] To be compelled is not merely to be *caused* to have some desire; rather it is to be caused to have a desire in a manner that bypasses the subject's capacities for control over their mental life.[48] Notice here that Mele appeals to objective facts about how the agent came to hold the desire in question; it can thus be construed as an objective historical account. Notice also that the account is framed negatively; authenticity requires the *absence* of certain types of causes. A further necessary condition is that the agent neither performed nor arranged for the bypassing that led her to develop the psychological characteristic in question.[49]

I shall have cause to refer to a further necessary condition that Mele specifies in the course of refining his view below. However, the first two conditions specified here are enough to see that Mele's account can offer a way of explaining how Beth might lack autonomy in a way that Ann does not. She is compelled to now value the life of a productive philosopher in a manner that bypasses her control over her mental life. The main challenge facing Mele's account as I have so far specified it is that it sets a seemingly high bar for autonomy. As those who espouse relational views of autonomy point out, we are all, at least in part, an outcome of social and environmental forces that determine many of our values and desires at a pre-critical stage of our development. There is '... no self before the socialization that creates it'[50] in pre-critical childhood development. In a sense then, by Mele's lights, we are all victims of manipulative processes that serve to undermine our autonomy, in so far as we have all had certain values and desires imputed to us during the pre-critical stages of our development, some of which we are now practically unable to shed. Accordingly, it seems plausible that autonomy is compatible with the fact that many of our desires were caused in ways that bypassed our mental control.[51]

Consider now John Christman's alternative subjective historical account. Rather than appealing to objective facts about how a desire was caused, subjective accounts instead ask '... if the person would have, or did resist the adoption of a value or desire, and for what reasons'.[52] In the early iteration of the view, Christman argued that the relevant question for autonomy is whether the agent would have resisted *the process* by which she came to have a particular desire (in a minimally rational, and self-aware manner). One obvious problem with this initial iteration of the view is a phenomenon I explored in the previous section. One can reject the process by which

[46] Ibid., 145. [47] Ibid., 166. [48] Ibid., 171. [49] Ibid., 166.

[50] Noggle, 'Autonomy and the Paradox of Self-Creation', 104.

[51] Christman raises a similar criticism at Christman, *The Politics of Persons*, 141.

[52] Christman, 'Autonomy and Personal History', 10.

one acquired a desire, and yet still hold that desire autonomously if one endorses it on other grounds. Partly in view of this objection, Christman has recently revised his subjectivist view by defending the following three necessary[53] authenticity conditions of an agent's being autonomous with respect to some basic evaluative characteristic C:

1. Were the person to engage in sustained critical reflection on C over a variety of conditions in the light of the historical processes (adequately described) that gave rise to C;
2. She would not be alienated from C in the sense of feeling and judging that C cannot be sustained as part of an acceptable autobiographical narrative organized by her diachronic practical identity;
3. The reflection being imagined is not constrained by reflection-distorting factors.[54]

Notice that condition 1 shifts the focus of the relevant reflection from the *causal history* of a particular desire, to the desire itself, *in light of* its causal history. This move circumvents a significant part of the above criticism. However, it does so at the cost of considerably weakening the relevance of history per se to authenticity. The relevant reflection now concerns the nature of the psychological characteristics themselves, rather than the manner in which one came to acquire them. Indeed, I suggest that the need for historical theories to make this move suggests that historical views of autonomy are in fact focusing on the wrong aspect of our desires, since these revised versions maintain that it is not the causal history of our desires that really matters with regards to our autonomy; what really matters is whether the agent *now* believes that they ought to endorse their desires. The history of the desire is one thing that may contribute to that decision, but it is not the only consideration.

This latter point emphasizes the fact that it is uncharitable to characterize the rationalist view that I have defended as entirely ahistorical. The rationalist view rejects what David Zimmerman refers to as source historicism; that is, it rejects the thesis that our autonomy with respect to a particular psychological property depends on the manner in which it is acquired. However, it is perfectly compatible with what Zimmerman calls process-historicism, that is, the thesis that 'autonomy grounding psychological states and processes are temporally extended'.[55] As I explained in the previous chapter, the constituents of our character systems have authority to speak for the 'true self' because they are diachronically extended, and relatively stable features of our psychological economies.[56] Process-historicism matters for autonomy, but it is not at all clear that source-historicism does.

We may also note that the objective historical account can also make a similar move to the one discussed above to circumvent Berofsky's concern. That is, the

[53] Necessary but not sufficient. These authenticity conditions are supplemented with three conditions concerning the competencies that are causally necessary for autonomy. See Christman, *The Politics of Persons*, 155.

[54] Ibid., 155.

[55] Zimmerman, 'That Was Then, This Is Now', 642.

[56] Indeed, what I refer to as the agent's character system shares a number of salient similarities with what Christman refers to as the agent's diachronic practical identity in Christman, *The Politics of Persons*.

objectivist can (and should) claim that an agent can initially acquire a desire in a manner that bypasses her mental control, and yet still be autonomous with respect to it, as long as she later exerts mental control by deciding to sustain that desire once she is made aware of its dubious causal history.[57] Indeed, Mele supplements his theory with the following necessary condition that responds to this kind of problem: S will only fail to be autonomous with respect to a particular value P which she was compelled to have in a manner she did not arrange if it is also true that:

> *S* neither presently possesses nor earlier possessed pro-attitudes that would support his identifying with *P*, with the exception of pro-attitudes that are themselves practically unsheddable products of unsolicited bypassing; then *S* is compelled* to possess *P*.[58]

With this condition, Mele's account moves closer to the view of manipulation that I defended in the previous section. In typical cases of manipulation where the agent endorses the changes evinced, it seems plausible that they do so by virtue of the fact that the new characteristic coheres with *pre-existing* elements of their character system. Such agents can be autonomous because they do not meet the above necessary condition of compulsion. Mele and I are in agreement on this point.

However, our approaches come apart when we consider cases in which the agent's endorsement of a manipulated psychological characteristic is *itself* a product of elements of one's character system that one has also been compelled to have. Crucially, Beth endorses her manipulated values in this kind of way. For Mele, endorsement of a manipulated value by *other* compelled values would meet the further necessary condition just outlined, and so such an agent would fail to be autonomous because they would meet all the necessary conditions of having been compelled to have the relevant values in a manner that she did not arrange. This is an important point, because it is here where the relational objection to the objectivist externalist account shows its teeth. Why does Beth lack autonomy in a way that most people do not, if their character systems as a whole are unsheddable in a relevantly similar way, by virtue of their formation in pre-critical periods of their lives?

One response to this problem is to appeal to the subjectivist approach, and claim that Beth lacks autonomy because she would hypothetically feel alienated from her new values were she to reflect on them in light of their causal history. I am not convinced by the subjectivist explanation of why Beth lacks autonomy, but we can put that point to one side.[59] Instead of attacking this explanation, I believe that we should adopt the revisionist view that *both* Beth and Ann are autonomous in an important sense, even if Beth meets all of Mele's conditions of compulsion in a way

[57] Mele, *Autonomous Agents*, 165.

[58] Ibid., 172.

[59] Briefly, my concern is that Beth could plausibly have been manipulated in such a manner that she would *not* feel so alienated. Christman's third condition outlined above is intended to block off this kind of example. Yet, it is not clear that it can be successful; for it to be so, we would need a good theory of what it is for a factor to be reflection-distorting in the relevant sense. In his discussion, Christman relies on our 'independent knowledge' of such factors. Yet, in borderline cases, this is precisely what seems to be missing. Consider for example a young woman who prioritizes the avoidance of weight-gain over all other values including her own survival; alone, the mere fact that we might diagnose such an individual as having a psychiatric disease tells us little about whether we should understand her mode of reflection to be distorted in the relevant sense.

that Ann does not. On this revisionist view, Mele's example is still powerful because it raises a plausible question about whether we should employ historical conditions on Beth's prospective *morally responsibility* following global manipulation. However, the point is that we should be wary of assuming that these intuitions translate straightforwardly to the claim that Beth is not autonomous.[60]

This is particularly true when we think about the way in which we understand autonomy in the biomedical sphere. To illustrate, suppose that following the manipulation, Beth is told that she has a medical condition that will result in paralysis unless she undergoes a neurosurgical procedure that is likely to cause a mild cognitive impairment (equivalent let us say to her losing 5 IQ points). Prior to her global manipulation, it may be that Beth would have prioritized her motor capacities far above a small reduction in her cognitive capacities. Following the manipulation though, let us suppose that her priorities have changed; she no longer cares for non-philosophical pursuits, and even a small reduction in her cognitive capacities would be hugely damaging. Here is the key point: all other things being equal, 'post-manipulation' Beth can clearly autonomously decide to refuse to consent to the procedure, even though she did not arrange for the global change in values that manipulation evinced, and which now grounds the autonomy of her decision. We may also note that a significant benefit of claiming that Beth is autonomous is that it obviates the problem facing Mele's theory, of how to explain the way in which we can generally be autonomous in a way that Beth is not, if all of our characters are grounded by values that appear to be unsheddable by his lights.

However, why should we think that the intuitive appeal of Mele's example is grounded in the fact that our judgements regarding Beth's autonomy and moral responsibility can diverge?[61] In defending a similar view, Nomy Arpaly alludes to the way in which our judgements about moral responsibility may be muddied by conflicting understandings of the notion of personal identity. However, for this argument to succeed, one would need to explain why these intuitions do not similarly affect our judgements about autonomy. An alternative explanation for why our judgements about autonomy and responsibility might differ can be sourced in Gary Watson's distinction between accountability and attributability. According to Watson, it can be possible for conduct to be *attributable* to an individual, where the conduct itself admits of appraisal and when it make sense to appraise the individual herself as an adopter of ends. Yet, the attributability of conduct does not entail that the agent is also accountable for that conduct, in the sense of her deserving sanction for it. One way in which we can cash out the conflicting claims about the responsibility and autonomy in the Beth/Ann case is to make the following two claims:[62] (i) these agents' conduct is attributable to them following global manipulation, but they

[60] Nomy Arpaly defends this revisionist view in Arpaly, *Unprincipled Virtue*, 126–30.

[61] Mele has offered a response to Arpaly's critique. However, he is mainly concerned with demonstrating that certain counterexamples raised by Arpaly to the bypassing condition (not provided here) fail. Crucially, he does not engage with the point regarding the putative differences between autonomy and moral responsibility, which are fundamental to Arplay's argument and my own criticism. See Mele, *Free Will and Luck*, 179–84.

[62] Watson, 'Two Faces of Responsibility', 263.

are not accountable for that behaviour and (ii) autonomy only requires that our conduct is attributable to us, whilst accountability may be necessary for some conceptions of moral responsibility.[63,64]

On the view that I am proposing here then, cases of global manipulation epitomized in the Beth/Ann case and identified in (ii)[c] in my schema above are primarily relevant to questions of personal identity and moral responsibility rather than autonomy, at least with regards to the sort of autonomy that is of practical interest in bioethics. Notably, although I have suggested Christman's subjective approach would imply that Beth lacks autonomy, it seems that the view could be amended to endorse the same conclusion on the Beth/Ann case that I have suggested here. To conclude my analysis of the role of history in decisional autonomy, I shall highlight two ways in which the theory I presented in the last chapter further departs from a subjective account that might be amended in this way.

As I discussed above, although considerations of history are no longer the primary consideration on Christman's revised theory, they still play a significant role; the agent must not (hypothetically) feel alienated from a given psychological characteristic *in light of its causal history*. However, whilst I agree that the dubious causal history of a desire may mean that we ought to critically assess the content of those desires to ensure that we endorse them, I remain sceptical of the claim that the agent's own attitude towards that causal history itself should matter with respect to the authenticity of their psychological characteristics themselves.

Indeed, an agent's own attitude towards the causal history of a characteristic can be irrelevant to their autonomy with respect to it. Suppose Alex loathes Ben and detests his world-view. Ben is giving a detailed, well-researched speech about why the government ought to adopt policy A rather than policy B. Although Alex previously endorsed policy B, he finds that he is rationally persuaded by the arguments in Ben's speech to now endorse policy A. Nonetheless, he feels alienated from this new preference, simply by virtue of the fact that it was Ben, his fiercely detested enemy, who succeeded in persuading him. By Christman's lights, it seems that Ben is not autonomous with respect to this new preference, but this seems implausible. The mere fact that Alex feels alienated from his preference because of Ben's role in it is not sufficient to show that Ben's rational persuasion is a form of controlling influence that serves to undermine Alex's autonomy.

Furthermore, Christman's theory can only provide practical guidance if we assume that we have an adequate grasp of the causal history of our psychological characteristics. However, the causal histories of some of our desires may remain opaque to us,[65] and we may even hold false beliefs about them. Indeed, this represents a considerable challenge in the psychiatric context, where clinicians may face the challenge of distinguishing between those forms of apparently psychotic phenomena

[63] For a defence of the claim that Beth/Ann are in fact both morally responsible, see Talbert, 'Implanted Desires, Self-Formation and Blame'.

[64] Of course, there may be other differences between the two concepts. For instance, we may lack autonomy due to reasons of ignorance without thereby lacking moral responsibility on the basis that our ignorance was culpable.

[65] Levy, *Hard Luck*, 105.

that constitute pathological but benign delusional states, from eccentric beliefs and values that are deeply embedded within an individual's character system.[66] In some cases, there may be little clear difference in the causal history of the cognitive states in question. Christman denies that his theory is problematic in this sort of way, because his account requires that the agent's conception of the relevant causal history need only be 'minimally adequate . . . in the sense that it be consistent with accepted evidence and known causal sequences'.[67] Yet, a recent application of Christman's theory in neuroethics reveals that the opacity of the causal histories of our psychological characteristic still affects the practical application of his theory.

In a recent paper, Daniel Sharp and David Wasserman have argued that Christman's theory can provide much needed illumination about questions of moral responsibility that are arising in problematic real-life cases in which patients who have undergone neurosurgical treatment (Deep Brain Stimulation) exhibit uncharacteristic behaviours following the intervention.[68] In considering a hypothetical example of an individual who develops a gambling addiction following neurosurgical treatment, and who endorses this new behaviour, the authors write:

> Many (ourselves included) have the intuition that the gambler is not fully responsible for his conduct because his endorsement itself *appears to be caused by personality-altering effects of DBS*.[69]

Setting aside the important point that Christman's theory is primarily intended as a theory of autonomy rather than responsibility, I want to focus on the emphasized phrase here. Philosophers and neuroethicists generally assume that changes to behavioural traits or personality that are sometimes observed following Deep Brain Stimulation are directly caused by the treatment itself. However, as Frederic Gilbert and colleagues have forcefully argued, we have very little evidence to suggest that this is the case; the phenomenon could also be adequately explained by a host of other mechanisms. It could be the result of the treatment unmasking extant psychiatric symptoms or of the patient experiencing difficulties with social integration following the amelioration of a chronic medical condition.[70]

The problem here is that the very context that the historical theory is apparently required to illuminate is one in which we lack even a minimally adequate understanding of the causal history of the relevant psychological characteristics. It is one thing to assume that an individual would hypothetically be alienated from a desire that has been induced by brain stimulation. It is another to assume that they would be alienated from a desire that they have formed as a result of their recovering from a chronic medical condition.

[66] For discussion of some enlightening case studies in this regard, and an argument for basing a distinction between spiritual and pathological forms of psychotic phenomena in considerations of their role in what I have called the agent's character system (and not the causal history of these cognitive states), see Fulford and Jackson, 'Spiritual Experience and Psychopathology'.

[67] Christman, *The Politics of Persons*, 154.

[68] Sharp and Wasserman, 'Deep Brain Stimulation, Historicism, and Moral Responsibility'.

[69] Ibid., 179, my emphasis.

[70] Gilbert, Viaña, and Ineichen, 'Deflating the "DBS Causes Personality Changes" Bubble'.

Finally, Christman stipulates that the reflection that his theory demands is only hypothetical, and that it can thus accommodate the thought that many central aspects of our lives have never been reflectively endorsed.[71] In contrast, rationalist theories tend to stipulate that critical reflection must be carried out at some point, even if only unconsciously, or in a dispositional sense.[72] The virtue of Christman's hypothetical approach is that it makes his account of autonomy significantly less demanding; however, in the biomedical context it also comes at a cost, since it renders the view difficult to operationalize. In assessing an agent's autonomy with respect to a decision, rather than simply enquiring about their general values, we not only have to know the causal history of the agent's desires, we have to make a judgement about what that agent would hypothetically feel about that causal history if it was brought to their attention.

This feature of the view raises the bar for third-party assessments of autonomy. However, in other ways the view also lowers the bar too far for the standards of decisional autonomy itself (rather than its third-party assessment); if it is true that an agent has *never* evaluated some central element of their psychological economy in any way, be it implicitly or unconsciously, then I do not hasten to conclude that the agent lacks autonomy with respect to that aspect of her psychology. Whilst this may seem a hard bullet to bite, we may note that those who endorse the hypothetical reflection condition have to bite the bullet of accepting that an agent qualifies as self-governing without ever actually attending to any element of his practical identity, no matter how minimally. I struggle to agree that this would be indicative of an agent engaged in any sort of active self-governance.

4. Informational Manipulation

At the beginning of this chapter, I distinguished psychological manipulation from informational manipulation, and deception. The latter two forms of influence affect the agent's beliefs, albeit in somewhat different ways. Beauchamp and Childress classify deception as a form of informational manipulation in outlining the standard view of autonomy. However, I believe that the clarity of the discussion will be best served by separating the two. One reason for this is that it can often be theoretically rational to believe *x* when one has been deceived into believing *x*, since we can be rationally justified in believing the testimony of others, even when it is false. In contrast, whilst informational manipulation may also involve leading another to develop false beliefs, it more typically involves leading an agent to adopt theoretically *irrational* beliefs.

We humans are subject to a considerable number of cognitive biases when it comes to forming our beliefs, and these biases can compromise our autonomy in a number of ways. Some biases may lead individuals to misapply their values, either by

[71] Christman, *The Politics of Persons*, 145.

[72] Ekstrom, 'A Coherence Theory of Autonomy'; Savulescu, 'Rational Desires and the Limitation of Life Sustaining Treatment'.

causing them to form a false belief about the world or to make basic logical errors; others can cause a patient to act in a way that does not reflect her values.[73]

Cognitive biases that can compromise autonomy in the first way include the phenomenon of motivated reasoning, in which agents regard an argument as fallacious simply because they are already predisposed to reject its conclusion.[74] Consider also the framing effect. Evidence from behavioural psychology suggests that if information provided to a patient is framed positively, then agents deciding on the basis of that information are more likely to be risk averse than if the information is framed negatively. Savulescu uses the following example to illustrate the importance of the framing effect in medical consultations:

> . . . (l)ung cancer can be treated by surgery or radiotherapy. Surgery is associated with greater immediate mortality (10 per cent v 0 per cent mortality), but better long-term prospects (66 per cent v 78 per cent five-year mortality). The attractiveness of surgery to patients is substantially greater when the choice between surgery and radiotherapy is framed in terms of the probability of living rather than the probability of dying.[75]

The framing effect engenders a form of theoretical irrationality because it involves logical incoherence; in the above example, it would be contradictory for a patient to prefer surgery when the comparative risk/benefit profiles are framed positively, but to also prefer radiotherapy when the (very same) comparative risk/benefit profiles are framed negatively.

Cognitive biases that can compromise autonomy in the second way include the bias that agents exhibit towards the present, and their reluctance to consider the possibility of future harms when they weigh their reasons for pursuing different outcomes.[76] For example, a patient may reject their physician's recommendation that they stop smoking, not because they believe that the pleasure they get from smoking is more valuable than increasing the probability of a longer lifespan, but rather because they fail to attend to the disvalue of the later consequences of smoking.

The evidence regarding cognitive biases suggests a further reason to reject value-neutral approaches to the shared decision-making model that I considered in section 1. It is a mistake to assume that the information we might need to give to individuals to ensure they adequately understand their options can be provided in a value-neutral way. Although the *content* of the information we provide can enable autonomy by ensuring adequate understanding, the *manner* in which it is presented can covertly influence individuals to develop irrational beliefs, or to act in practically irrational ways. Indeed, it seems that some nudge techniques are designed to capitalize on the forms of theoretical irrationality to which our propensity to cognitive biases makes us particularly vulnerable.

However, it is not always clear how we should demarcate kinds of informational manipulation from forms of influence that enhance the understanding that is necessary for decisional autonomy. To illustrate, Blumenthal-Barby and Burroughs suggest that the use of vivid examples and explanations can constitute a nudge,

[73] Levy draws this distinction in Levy, 'Forced to Be Free?', 298.

[74] Ibid., 298.

[75] Savulscu, 'Rational Non-Interventional Paternalism,' 328–9. See also Brock, *Life and Death*, 88.

[76] Levy, 'Forced to Be Free?', 6. See also Brock, *Life and Death*, 84.

noting that these items can elicit strong emotional responses that powerfully shape decisions and behaviours. By way of example, they describe the following study by Volandes et al.:

> A group of elderly adults was shown a 2-minute video about what life with advanced dementia was like along with a written description, while the other group was just given the written description. The group that saw the video had 86% of its members indicate that they would want 'comfort care only' in such a state, whereas in the control group that number was only 64%.[77]

Blumenthal-Barby and Burroughs' discussion suggests that they understand this to be an instance of the use of affect to influence decision-making, and that such use of affect will typically amount to a manipulative interference.[78] However, it seems that more needs to be said in favour of this interpretation, particularly in light of my discussion of emotions and persuasion in section 2. In particular, it seems plausible that the video in question may simply have made reason-giving facts about the badness of life with dementia more vivid to the viewers; if so, it seems that the strategy should be construed as facilitating rather than impeding rational decision-making.[79]

In light of my discussion of the kinds of psychological and informational manipulation that nudge techniques can employ, what should we say about the implications of these techniques for decisional autonomy? Advocates of these techniques often argue that they are compatible with individual autonomy by appealing to the fact that they do not limit the agent's choice set (unlike bans) or involve significantly altering incentives.[80] Whilst true, such observations miss what is primarily at stake in the debate about nudging and autonomy. The fact that nudges do not undermine autonomy in some ways (by restricting freedom or coercing) does nothing to answer the fact that they may yet pose other threats to our decisional autonomy by inducing forms of theoretical and/or practical irrationality.[81]

Proponents of nudges alternatively might advert to the fact that normal human decision-making is plagued by non-rational influences.[82] Here, it might be claimed that nudges can do little to undermine decisional autonomy if we *already* lack such autonomy. Of course, the fact that non-rational influence of some sort is inevitable in a given choice domain does not imply that steps should not be taken to mitigate these non-rational effects. Recall the example of the framing effect above. In this circumstance, even though the physician has to make a choice about whether to frame the

[77] Blumenthal-Barby and Burroughs, 'Seeking Better Health Care Outcomes'.

[78] Ibid., 5.

[79] Indeed, these videos might prompt the sort of 'vivid imagination of alternatives' that Savulescu argues is a requirement of autonomous decision-making. See Savulescu, 'Rational Desires and the Limitation of Life Sustaining Treatment'.

[80] Thaler and Sunstein, *Nudge*.

[81] Saghai has developed a philosophically robust defence of a closely related argument that nudges are compatible with autonomy, by invoking considerations of freedom of choice and resistibility. See Saghai, 'Salvaging the Concept of Nudge'. See Ploug and Holm, 'Doctors, Patients, and Nudging in the Clinical Context—Four Views on Nudging and Informed Consent' for a rebuttal of this argument drawing on rationalist themes.

[82] Thaler and Sunstein, *Nudge*; Blumenthal-Barby and Naik, 'In Defense of Nudge–Autonomy Compatibility'.

information positively or negatively, this does not entail that he cannot seek to then mitigate the non-rational influence his framing might have on the patient's choice, by asking the patient to justify or explain their choice.[83]

Notwithstanding these claims, I shall suggest below that nudges may infringe an interpersonal form of voluntariness that standard non-rational decision-making does not. To conclude this part of the discussion though, we should, I believe, concede that some (but not all) of the nudge strategies that aim to influence individual behaviour are manipulative in a manner that serves to undermine autonomous decision-making. Despite the powerful, and potentially beneficial effects of such strategies, we should not labour under the illusion that interventions are always compatible with local autonomous choice. As I argued in section 2, they may fail to be so if they bypass and subvert the cognitive element of our practical rationality. However, they will also fail to be so if they engender forms of theoretical irrationality.

Again, it is important to be clear that this is a point about the implications of nudges for decisional autonomy, and not an all things considered moral judgement on their use. One might argue that broadly beneficence-based concerns could outweigh these considerations of local autonomy; however, such a strategy will naturally require that one is able to respond to the (justified) allegation of paternalism that would be weighed against it.[84] An alternative, and I believe more promising strategy, might seek to justify these strategies of influencing behaviour by appealing to the value of the agent's global, rather than local autonomy. I shall explore this point when I consider the value of different kinds of autonomy in Chapter 9.

5. Deception

On one prominent understanding, deception involves imparting false beliefs to another person.[85] If it is the case that decisional autonomy requires that agents hold certain decisionally necessary true beliefs, then deception (so-construed) will serve to undermine autonomy just in so far as it leads individuals to hold false beliefs about features of their choice domain that are subjects of what I am terming 'decisionally necessary' beliefs. However, the account of autonomy and true beliefs that I have been sketching so far (and which I shall flesh out further in Chapter 5) points towards a broader account of deception. I have suggested that decisional autonomy can require that agents hold certain true beliefs about features of the decision in question; crucially, it seems that there can be cases in which one causes another to fail to hold the relevant true beliefs by *omitting* key information. If this is

[83] Ploug and Holm, 'Doctors, Patients, and Nudging in the Clinical Context—Four Views on Nudging and Informed Consent'. See also Miller and Fagley, 'The Effects of Framing, Problem Variations, and Providing Rationale on Choice'. In their discussion, Ploug and Holm imply that the framing effect undermines autonomy in so far as it undermines the understanding required for autonomous decision-making. I agree with Blumenthal-Barby and Naik's criticism of this claim. Blumenthal-Barby and Naik, 'In Defense of Nudge–Autonomy Compatibility'.

[84] This strategy corresponds to what Ploug and Holm describe as the 'priority view'. See Ploug and Holm, 'Doctors, Patients, and Nudging in the Clinical Context—Four Views on Nudging and Informed Consent', 36 for discussion.

[85] Shiffrin, *Speech Matters*, 19.

so, then it is also possible to undermine autonomy by deception via omission on the theory that I am outlining here.[86] Accordingly, in the following discussion, I shall seek to defend a broader conception of deception as involving either causing another to hold false beliefs, or causing them to fail to hold decisionally necessary true beliefs.

This is a controversial way of broadening the scope of deception. Another controversial implication of the view that I have so far defended is that deception need not be intentional. In discussing psychological manipulation I noted that the claim that manipulation need not be intentional follows straightforwardly from the claim that rational authenticity is a necessary condition of decisional autonomy. Strikingly, an analogous claim can be made with regard to deception if one holds that sufficient understanding is a necessary condition of autonomy. Deception need not be intentional in order to undermine decisional autonomy as long as it serves to lead agents to fail to hold decisionally necessary beliefs.

The clarity of the following discussion will be aided by making some distinctions between possible forms of deception. Of course, one of the most common methods of deception is lying. However, lying is not co-extensive with deception, in so far as a lie does not entail that a liar successfully imparts a false belief in the manner that deception connotes.[87] We may say that an agent lies, when she intentionally provides her target with information that she believes to be incorrect, and her doing so manifests her intention to get her target to treat the information as an accurate representation of what she (the liar) believes.[88] A lie will also deceive if this has the effect of imparting a false belief to the target. In contrast, unintentional deception can occur when one provides an agent with information that they believe to be true, but which is in fact false. Deception via omission occurs when the target is led to develop a false belief, or to fail to hold a decisionally necessary true belief because of the omission of certain key information.

Of course, a great deal here turns on the feasibility of decisionally necessary beliefs. I briefly defended this view in the previous chapter, and I shall offer a more principled defence in Chapter 5. Here though, to illustrate deception via omission in a medical context, a physician may so deceive their patient by providing them with only a partial disclosure about their condition, or employing euphemisms to obscure the true nature of the condition. In such a case, deception via omission may lead the patient to explicitly hold false beliefs, or to fail to hold true beliefs of the sort that are crucial for making certain future decisions with the kind of understanding that autonomy requires. Suppose that tests revealed that Maurice has motor neurone disease; however, instead of explicitly informing Maurice of this particular diagnosis, the physician tells him that he has a condition that will cause him increasing weakness, but that he will be made 'as comfortable as possible'. On the basis of the

[86] For another detailed defence of this view, see Cox and Fritz, 'Should Non-Disclosures Be Considered as Morally Equivalent to Lies within the Doctor–Patient Relationship?'

[87] Shiffrin, *Speech Matters*, 19–21.

[88] Ibid., 13. Notice that on this account, it is possible to lie without deception in the sense that '. . . a lie does not depend on its recipient being deceived'. In such cases, I suggest that the lie may not undermine the voluntariness of the recipient's decision, but that there may nonetheless be reasons to sanction the liar. Another interesting feature of Shiffrin's account is that lying must be intentional in the sense indicated above, but it need not involve the intention to deceive.

conjunction of this euphemism, and the partial disclosure about the effects of the increasing weakness caused by motor neurone disease, Maurice forms the false belief that his condition is not all that serious. If so, on the definition that I am employing here, the physician would have deceived Maurice via omission, even if this were not his intention; Maurice fails to understand his situation in a manner that allows him to draw accurate connections between his values and his available options.

There is a significant philosophical debate as to whether lying is morally on a par with deception via omission.[89] However, as I mentioned at the outset of this chapter, I am interested only in the question of the effects of different sorts of influence on autonomy. Insofar as lying and omission can lead individuals to either form the same false beliefs or to fail to have decisionally necessary true ones, I claim that both can undermine autonomy. However, there are some important distinctions between the two. First, there is often a straightforward causal connection between the telling of a lie, and the target holding a false belief, such that the target can straightforwardly blame the liar for the fact that they hold a false belief. However, in the case of deception via omission, this causal relationship is far less straightforward, particularly when the deception is non-intentional. The explanation for this is that it seems plausible that autonomous agents have *some* degree of doxastic responsibility to obtain their own true beliefs about the world. If so, the fact that others omit to provide one with information cannot be said to be the only causal factor in one's ignorance. Just as we can distinguish between culpable and non-culpable ignorance in discussions of moral responsibility and blame, it also seems possible to distinguish between autonomy-undermining ignorance that is the fault of a third party, and that which is the fault of the agent herself.[90]

Accordingly, in some cases an individual's failure to hold a decisionally necessary belief is not best attributed to the fact that *others* have omitted to provide certain information. In everyday life, individuals plausibly have some responsibility to make attempts to understand the situations in which they find themselves, and the reasons that obtain for them in those situations. However, in biomedical decision-making, patients place a great deal of trust in their physician due to the considerable knowledge gap that exists between them with regards to salient medical facts. An upshot of this is that patients may transfer much of their everyday responsibility to cultivate decisionally necessary beliefs onto the physician in this context, in the form of a presumed duty of the physician to disclose information that is necessary for the patient to make an autonomous treatment decision. This represents an important way in which individuals in the biomedical context are more vulnerable to deception (broadly conceived) than they are in everyday life.

The understanding of deception I am employing here also runs contrary to the claim that only *intentional* deception undermines autonomy.[91] Interestingly,

[89] Pugh et al., 'Lay Attitudes toward Deception in Medicine'; Benn, 'Medicine, Lies and Deceptions'; Gillon, 'Is There an Important Moral Distinction for Medical Ethics between Lying and Other Forms of Deception?'; Jackson, 'Telling the Truth'; Bakhurst, 'On Lying and Deceiving'.

[90] For discussions of the doxastic responsibilities of patients, see Kukla, 'How Do Patients Know?'; Foster, *Choosing Life, Choosing Death*, 104.

[91] See Wilkinson, 'Nudging and Manipulation'.

Beauchamp and Childress claim that only intentional deception undermines autonomy. This is somewhat perplexing given their view that substantial understanding is a necessary condition of autonomy; why suppose that only intentional forms of deception can subvert substantial understanding? The claim that *only* intentional deception undermines autonomy is on more solid ground when it is held in conjunction with the claim that autonomy does not require certain true beliefs. Wilkinson, for instance, adopts this strategy.[92] He writes:

> A person may have false beliefs about his or her options without his or her autonomy being affected; who has true beliefs about all their options? But if those beliefs came about through deceit, his or her autonomy has been harmed.[93]

However, the fact that holding *some* false beliefs is compatible with decisional autonomy does not entail that *any* particular false belief about one's options is compatible with decisional autonomy. To illustrate, an individual can autonomously decide to undergo a medical procedure on the basis of a belief that it will be successful in ameliorating their condition, even if their belief turns out to be false. This reflects the fact that not all sorts of information about our choices are decisionally necessary, a point I raised in Chapter 2. However, this is quite compatible with the claim that some information is. To repeat an example from earlier in the book, an individual cannot be said to have autonomously decided to undergo a vasectomy if they fail to understand that it will lead them to be infertile.

More generally, simply showing that autonomy is compatible with individuals holding some false beliefs is not sufficient to demonstrate that decisional autonomy does not require that individuals must hold any true beliefs. To appeal to the compatibility of autonomy with certain false beliefs in order to deny the existence of *any* decisionally necessary beliefs is rather like pointing to a white swan in order to disprove the possibility of a black one. Following the Aristotelian claim that we can sometimes fail to be autonomous due to reasons of ignorance, and in accordance with the standard account's criterion of understanding, we should, I believe, acknowledge the possibility of decisionally necessary beliefs, and their implications for our understanding of deception. I shall offer a principled approach to identifying decisionally necessary beliefs in Chapter 5.

There are of course important non-autonomy based moral reasons to separate out the different forms that deception can take. Intentional deception plausibly involves wrongs that non-intentional deception does not, and intentional deceivers will often be culpable in a manner that may not be the case if deception was unintentional. Indeed, from a legal perspective, the question of whether a physician intentionally lied to their patient or unintentionally omitted vital information in obtaining consent to a medical intervention might make the difference between the procedure being an instance of battery rather than negligence. However, from the perspective of the

[92] For other examples of theorists who claim that false beliefs do not undermine autonomy, see McKenna, 'The Relationship between Autonomous and Morally Responsible Agency', 208–9; Arpaly, 'Responsibility, Applied Ethics, and Complex Autonomy Theories', 175.

[93] Wilkinson, 'Nudging and Manipulation'. Note that Wilkinson uses this observation to defend the view that only intentional manipulation undermines autonomy.

individual's autonomy *alone*, my claim is that failing to hold decisionally necessary true beliefs undermines an individual's ability to make an autonomous decision, no matter how they were influenced to fail in this way.

There are various ways in which a physician can either intentionally or unintentionally deceive a patient. Lying, of course is the most obvious method. However, there are also more subtle means of deception. For instance, as I illustrated with the example of Maurice above, the physician may not provide the patient with any false information, but simply be selective about the information that they choose to divulge to a patient, so that the latter forms an inaccurate impression of their condition, an impression that means that they do not fully understand the salience of the choices they face. These observations about deception via omission raise important questions about how we should understand the standards of information disclosure that valid consent requires in a medical context. I shall postpone this discussion until Chapter 6. To conclude this chapter though, I want to reconsider the role of intentionality in controlling influence, and the further interpersonal sense of voluntariness it connotes.

6. The Role of Intentions and Interpersonal Voluntariness

I have so far defended a view of controlling influences that downplays the necessity of third-party intentional agency in determining whether a particular form of influence undermines decisional autonomy. As I mentioned above, my approach in this regard is a corollary of the fact that I have endorsed (i) a rationalist authenticity condition on decisional autonomy and (ii) the possibility of decisionally necessary beliefs. However, I do not mean to claim that the interference of intentional agents on another's autonomous decision-making is therefore morally equivalent to non-intentional or non-agential interference. There are clearly some important differences between the two that are not primarily grounded in their implications for the sense of autonomy that I am outlining here. First, in the former case, the interfering agent violates the Kantian imperative that enjoins one to act in such a way that one treats other rational agents never merely as a means, whilst the same need not be true in the latter. Second, whether or not one agent intentionally interfered with another may be significant with regards to assessments of culpability for the harm caused through one's interference. In many cases, I believe that our concern with the intentions of those who exert influence over us is primarily grounded in moral concerns that are largely orthogonal to the question of the target's autonomy per se. Moreover, the ambivalence that some authors claim to have about whether a certain form of influence (such as manipulation)[94] necessarily requires intentional agency, may be attributable to the fact that we may be invoking the concept of the influence at stake to answer quite different moral questions.

Nonetheless, there does seem to be some intuitive plausibility to the claim that intentional interference is somehow worse from the point of view of the agent's

[94] Barnhill, 'What Is Manipulation?'

autonomy. The agent who is intentionally deceived seems to have experienced a greater affront to her autonomy than the agent who is unintentionally deceived. Despite the claims that I have so far advanced in this chapter, I believe that this intuition captures an important truth. Crucially though, and contrary to what some theorists have implicitly claimed, it is not the *only* truth about autonomy.

The Aristotelian distinction between two types of non-voluntary action (those performed from reasons of ignorance and those that take place by force) captures aspects that are central to our understanding of decisional autonomy, at least in the biomedical sphere. In this concluding section though, I want to consider the possibility that when an agent's deficit in decisional autonomy is attributable to intentional third-party agency, this can be understood to undermine a separate kind of 'freedom from domination' that undergirds a sense of voluntariness that the Aristotelian distinction fails to acknowledge. I also suspect that this further form of voluntariness also adds further fuel to the fire regarding our ambivalence about the necessity of intention to our concepts of different forms of controlling influence.

The conception of freedom I have in mind here might be understood to denote a particular kind of negative freedom, namely the absence of positive constraints that have been intentionally imposed *by another agent*. This freedom is a specific kind of the broadly libertarian type of freedom that stresses the importance of the absence of constraints. However, freedom from domination has historically been understood in a broader sense within the republican tradition.[95] In this tradition, it is noted that dominance over another can be exerted even if it does not involve the *actual* imposition of constraints. For example, if another agent has the mere *capacity* to arbitrarily interfere with another's choices, they may be said to dominate the other in a sense that undermines the latter's freedom from domination.[96]

Accordingly, on the republican understanding, this freedom is violated if another has the mere capacity to interfere with one's choices. In contrast, on the libertarian understanding of freedom from domination as a particular kind of positive constraint, this freedom is only undermined if the dominating party does *in fact* exercise that capacity and actively interferes with the agent's decision. The sense that I mean to invoke here is the libertarian conception, although this is not the place to try and settle the debate as to whether it is more plausible than the republican conception. Although very little of what I shall claim turns on the fact that I endorse the libertarian rather than republican understanding of this freedom, it is important to acknowledge that the libertarian conception of this freedom is more robust. In order to undermine it, one must actually interfere with another's choices; it is not sufficient to merely have the capacity to do so.

[95] Pettit, *Republicanism*; Pettit, 'Freedom as Antipower'; Skinner, *Liberty before Liberalism*. For a discussion of the concept of domination itself, see Lovett, 'Domination'.

[96] This republican understanding of freedom can thus be invoked by those who claim that Savulescu and Persson's famous God Machine example involves the violation of individual freedom, even for those law-abiding individuals whom the machine does not directly act upon. See Savulescu and Persson, 'Moral Enhancement, Freedom, and the God Machine'; Sparrow, 'Better Living Through Chemistry?' for a republican response.

The key element of freedom from domination for my purpose here is that it is an *interpersonal* form of freedom; a lack of this freedom amounts to the subjugation of one's own will to another's authority. It may thus plausibly be construed as referring to a different sense of voluntariness than the one that is captured by the cognitive or reflective elements of decisional autonomy. The sense of voluntariness that reflective autonomy captures is the sense that is grounded by the thought that voluntary choices must reflect the agent's own character. The sense of voluntariness that the cognitive element captures is the sense that is grounded by the thought that ignorance can preclude us from acting effectively in pursuit of our ends, by rupturing the connection between our beliefs about our choices and our values. In contrast, the sense of voluntariness that freedom from domination captures is the sense that is grounded by the claim that it must be the agent *herself*, rather than other parties, who is in control of her decision-making if she is to be autonomous. This is the thought that Robert Wolff seeks to capture in his claim that 'The autonomous man, insofar as he is autonomous, is not subject to the will of another'.[97]

Why should we suppose that freedom from domination matters? Part of the explanation might be phenomenological; perhaps it is simply the case that third-party interference feels like more of an affront to our autonomy.[98] Elinor Mason, however, goes deeper than this phenomenological point in her discussion of this difference:

> What agents do to us is different to what non-agents do to us . . . when there is another agent, that agent takes the place of 'self' in self-determination. Other agents are qualified to do that because they themselves have wills and are self-determining – the blind forces of nature cannot take over in the same way.[99]

A claim that seems implicit in Mason's comment here is that there is a difference in *the nature* of the lack of control of an agent whose decision-making is subjected to intentional interference, and an agent whose decisional autonomy is undermined by forces of hazard.

The difference can be helpfully illustrated by way of analogy. The sense of voluntariness captured by the two elements of decisional autonomy I have discussed prior to this point may be understood as pertaining to the strength of a ship captain's grip on her vessel's helm, and her ability to navigate to her destination. The captain has the relevant control to the extent that she (i) has true beliefs about where to go, and (ii) is able to dictate the ship's movements through her own influence on the helm. She may lack control because she is lost, or because the wheel is simply left spinning, and the course of the ship is left to the uncontrolled dictates of the sea and wind. In contrast, the interpersonal sense of voluntariness may be understood as pertaining to whether it is the captain, or a usurper who has taken control of the helm; there is an important difference between the course of one's ship being left to the vagaries of the elements, which have little interest in your destination, and your

[97] Wolff, *In Defense of Anarchism*.

[98] Wertheimer, 'Voluntary Consent', 244–5.

[99] Mason, 'Coercion and Integrity', 196. Jennifer Blumenthal-Barby similarly denies the moral equivalence of environmental and agential influences on autonomy in Blumenthal-Barby, 'A Framework for Assessing the Moral Status of "Manipulation"', 125–6.

course being decided upon by an intentional agent who takes great interest in where you end up.

As I mentioned in the introductory chapter, some theorists claim that only intentional forms of controlling influence undermine autonomy. Such theorists place a great deal of stock in the sense of voluntariness that I am outlining here. In contrast, on the account of autonomy and controlling influence that I am outlining here, this sense of voluntariness supplements those outlined in the Aristotelian distinction. Decisional autonomy can be undermined by either non-agential processes (such as psychiatric disease) or the intentional interference of third parties. However, in the latter case, the influence exerted may serve to nullify an additional interpersonal sense of voluntariness. Crucially though, on the account of decisional autonomy that I have developed, one can lack decisional autonomy even if one's freedom from domination has not been violated. This distinguishes my account from those theories that claim that *only* intentional agents can undermine another's autonomy.[100] Indeed, I suggest that in order for one's will to have been dominated by intentional manipulation or deception, it must be the case that the interference in question has undermined the reflective or cognitive element of one's decisional autonomy. If the intentional deception or manipulation does not succeed in undermining one's decisional autonomy in a particular instance, it is difficult to make sense of the claim that one's will has been dominated in any significant sense by those exerting the influence.

One of the trends in the philosophical literature has been to understand the senses of voluntariness incorporated into decisional autonomy as being a matter of *solely* reflective autonomy, or *solely* freedom from domination. Both of these views, however, are mistaken. Our beliefs matter for decisional autonomy, and *both* agential and non-agential influences on our behaviour can undermine our autonomy. To return to the above analogy, one can fail to be in control of one's ship either because one does not have a strong enough grip on the helm, or because another has usurped one's position at the helm. Yet the former, which we may term *non-autonomy*, is not equivalent to the latter, which we may term *heteronomy*; this is not just so from the perspective of morality all things considered. It is also true from the perspective of interpersonal voluntariness.

Conclusion

In this chapter, I have outlined an approach to understanding different forms of manipulation and deception in light of the rationalist conception of autonomy that I developed in Chapter 2. In doing so, I also highlighted the significance of an interpersonal form of voluntariness not captured by the Aristotelian distinction.

I did not, however, address a salient form of controlling interference in this chapter, namely coercion. The reason for this is that coercion admits of greater

[100] Bublitz and Merkel, 'Autonomy and Authenticity of Enhanced Personality Traits'; Taylor, *Practical Autonomy and Bioethics.*

theoretical complexity than deception and manipulation. Moreover, as I shall explore in the next chapter, acknowledging the sense of freedom from domination that I explored in the second half of this chapter is crucial to providing a plausible theoretical basis for the ambiguous effects that coercion seems to have with respect to the voluntariness of the choices made in coercive situations.