2

Rationality and Decisional Autonomy

With the preceding chapter's discussion in mind, I am now in a position to consider what role rationality might play in decisional autonomy. Recall that the standard account of autonomy in bioethics claims that decisions are only autonomous if they are made intentionally, with understanding, and in the absence of controlling influences. However, as I pointed out in the introduction, there are some cases in which our intuitions speak strongly in favour of the claim that an agent can lack autonomy with respect to their decision, even though it meets the conditions set out in the standard account. Moreover, the standard account lacks a deep explanation of what constitutes a controlling influence.

Recall Jane, the unwilling addict who acts on a compulsive desire to take drugs. If Jane's failure of autonomy here could be attributed to her being irrational in some sense, then this would provide some motivation for claiming that the standard account should be supplemented with a rationality condition that precludes these agents from being autonomous with respect to irrational decisions.[1] However, this strategy raises three important questions. First, we might ask whether *all* forms of irrationality preclude autonomous choice. Second, we might ask whether the rationality of a decision makes a positive contribution to an agent's autonomy with respect to it, or whether we should simply make the weaker negative claim that irrationality precludes autonomy. Finally, and most importantly, we might wonder whether we can say anything more to justify the general strategy of appealing to rationality conditions to supplement the standard account, other than the fact that it accords with our intuitions in certain paradigm cases.

I shall answer these questions in this chapter by outlining an account of the role that theoretical and practical rationality play in decisional autonomy. In doing so, I shall particularly contrast my view with Rebecca Walker's recent defence of a rationalist account of autonomy. Walker endorses a negative rationality criterion on autonomy, according to which both practical and theoretical irrationality preclude autonomous choice. However, she does not commit herself to the claim that rationality might positively contribute to the autonomy of a decision (although she does leave open that possibility).[2] Instead, she claims that the 'straightforward' explanation for why one cannot be autonomous with respect to an irrational choice is that '. . . choosing irrationally is choosing on the basis of an error'.[3]

[1] Walker, 'Respect for Rational Autonomy', 343. [2] Ibid., 344. [3] Ibid., 344.

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

Whilst I agree with Walker about the significance of both theoretical and practical rationality, I shall argue that we need a deeper explanation of the role that rationality plays in autonomy than she provides. In outlining my own account of this, I shall suggest that a deeper explanation points us towards the view that rationality makes a positive contribution to autonomy. I shall begin by explaining why autonomous decision-making requires some degree of theoretical rationality, before turning to consider practical rationality.

1. Theoretical Rationality and Autonomy

In the introduction, I claimed that the standard account of autonomy reflects Aristotle's distinction between two types of non-voluntary action. In particular, I suggested that the criterion of understanding in the standard account reflects Aristotle's claim that an action is non-voluntary if it is performed from reasons of ignorance. Understanding is thus crucial to our ability to make voluntary choices in this sense; as Savulescu and Momeyer rightly point out, 'we cannot form an idea of what we want without knowing what the options on offer are like'.[4] We may add to this that in some cases a person may fail to understand the significance of their choice because they do not understand certain key features of their alternatives.

The criterion of understanding thus implies that agents must hold at least some *true* (and not merely rational) beliefs about their alternatives if they are to make an autonomous decision in that particular choice context. Call these 'decisionally necessary' true beliefs. What sort of beliefs might qualify as decisionally necessary? This is a complex question that I shall only be able to answer once further theoretical claims are in place (in Chapter 5). Roughly here though, we may say that there are at least some true beliefs that an agent must hold if they are to be able to minimally draw accurate connections between their values and their available options, in the manner that autonomous decision-making seems to require. Crucially, this view does not entail the strong claim that autonomous decision-making requires that we *only* choose on the basis of true beliefs; this is implausibly strong, given that we often cannot know for certain whether our beliefs are true. This is most clearly the case with our beliefs about future states of events. However, this does not mean that there cannot be *any* true beliefs that an agent must hold in order to make an autonomous decision.

To give one example here, suppose that a patient decided to undergo a vasectomy without understanding that this procedure will render him infertile. It seems doubtful that such a decision could qualify as autonomous. The individual has no idea about the implications that the procedure will have for him; we can even go further and say that it is doubtful that the patient in this case is even consenting to a vasectomy at all if he lacks this understanding. In my view, the belief that a vasectomy will cause infertility is thus decisionally necessary; to make an autonomous decision, we must know what our options are like in some minimal sense. This is a corollary of

[4] Savulescu and Momeyer, 'Should Informed Consent Be Based on Rational Beliefs?', 283.

the Aristotelian claim that actions performed from reasons of ignorance are in an important sense non-voluntary.

I shall attempt to further flesh out the concept of decisionally necessary beliefs later in the book. Here though, I am interested in the point that the criterion of adequate understanding may not preclude all forms of ignorance that are inimical to decisional autonomy. One's decision may be grounded in reasons of ignorance even when one holds the relevant decisionally necessary true beliefs. The explanation for this is that an individual may be theoretically irrational with respect to the way in which they *use* this information.

One illustration of this is the example of the patient I raised in my initial discussion of theoretical rationality in the previous chapter, who infers that they are likely to die from a surgery involving anaesthesia. Indeed, Savulescu and Momeyer raised this case to defend the view that autonomy requires theoretical rationality.[5] To illustrate the point further, Walker provides an example of a woman, called Maureen, who has been diagnosed with HIV/AIDS, and who refuses medication that there is a strong evidence base to suggest will be statistically likely to dramatically increase her chances of survival. However, Maureen believes that her statistical chances of survival with and without treatment are irrelevant to her, simply because they do not affect the more basic fact of fate that it is either the case that she will die in the next ten years, or she will not. In short, although she understands the relevant information and the statistical evidence about the treatment, her other fatalistic belief prevents her from applying this evidence to her own case.[6]

One might argue that the individuals in both of these cases of theoretical irrationality would fail to qualify as autonomous even on the standard account of autonomy, because they do not truly understand the information relevant to their decision if they reason in these ways. I am not convinced that the standard account's criterion of understanding is intended to capture such forms of theoretical irrationality, but the point is somewhat moot for my purposes here.[7] The reason for this is that if an advocate of the standard account conceptualizes the understanding criterion in this way, then this amounts to the concession that autonomous choice is precluded by irrational beliefs.

So why should we think that theoretical irrationality undermines autonomy? Is it simply the case that theoretical irrationality only undermines autonomy, or can theoretical rationality make a positive contribution to decisional autonomy? Walker advocates the former view, and justifies this by adverting to the further claim that theoretical irrationality undermines autonomy because it entails that one chooses on the basis of an error. Of course, this will only be a satisfactory explanation if *all* errors undermine the autonomy of the choices to which they lead. Yet this seems unlikely; indeed, Walker herself denies this, since she denies that true beliefs are necessary for autonomy.[8] By her own lights, autonomy is compatible with choosing on the basis of *some* errors in belief, namely decisions based on rational but false beliefs. Moreover, as I noted in the previous chapter, failures of theoretical

[5] Savulescu and Momeyer, 'Should Informed Consent Be Based on Rational Beliefs?'

[6] Walker, 'Respect for Rational Autonomy', 347.

[7] For exegesis of the standard account on this point, see ibid., 349.

[8] Ibid., note 11.

rationality can be compatible with true beliefs, as was the case in the example of the Othello syndrome; why should we suppose this error undermines autonomous decision-making?

We might also observe that Walker's denial of the importance of true beliefs is somewhat in tension with her apparent endorsement of the standard account's criterion of understanding. As I suggested above, the criterion of understanding implies that some true beliefs may be necessary for decisional autonomy. I shall further support this view later in the book, but we can leave that support aside for the time being. The point I wish to make here is that if Walker's negative approach is to be convincing, then it needs to be supplemented with a deeper explanation of why she thinks *all* errors of theoretical rationality threaten autonomy (even in cases where they do not lead to false beliefs), an explanation which is also compatible with her commitment to the claim that autonomy is compatible with holding (rational) false beliefs.

Alternatively, one can endorse a different view of the relationship between theoretical rationality and autonomy, one which is compatible with the thought that autonomous decision-making requires that the individual holds at least *some* true beliefs. On this view, we should avoid the claim that autonomous choice is compatible with choosing on the basis of *only* false beliefs, or with *complete* ignorance about information that is crucial to one's choice, as Walker seems to imply. This view is implausibly strong, if there can be decisionally necessary beliefs.

The claim that an individual must hold some true beliefs in order to be autonomous with respect to a particular decision is implicitly defended by Julian Savulescu. Savulescu argues that a necessary condition of autonomy is that individuals make their decisions on the basis of rational desires. In turn he defines a rational desire as one that results from an evaluation of the alternatives available, according to which one option (say A) is better than the other (B). The evaluation must involve at least the following three elements:

(1) knowledge of relevant, available information concerning each of the states of affairs A and B,
(2) no relevant, correctable errors of logic in evaluating that information, and
(3) vivid imagination by P of what each state of affairs would be like for P.[9]

I agree with the spirit if not the precise letter of Savulescu's view. In view of the way in which I distinguished practical and theoretical rationality in the previous chapter, I am reluctant to claim that these are conditions of rational 'desires'. Instead, I suggest that condition (1) (and to some extent [3]) pertains to the kinds of true (and not merely rational) beliefs that individuals must have in order to make a locally autonomous decision. Condition (2) in contrast is a theoretical rationality condition on autonomy. However, we may also note that the language of 'evaluation' that Savulescu employs suggests that considerations of practical rationality also have an important role to play in his view, as I shall explore below.

[9] Savulescu, 'Rational Desires and the Limitation of Life Sustaining Treatment'.

On the alternative view that I am outlining here, theoretical rationality can positively contribute to autonomy because when we form and sustain our beliefs in a theoretically rational manner, our beliefs are more likely to be true. In some cases our decisional autonomy may be *enabled* by our holding certain true beliefs, if these beliefs are decisionally necessary. I shall attempt to offer a deeper explanation of why true beliefs matter for decisional autonomy in this way in Chapters 5 and 6.

We can also make a stronger claim about the relationship between theoretical rationality and autonomy. Abiding by the norms of theoretical rationality can be important not just because doing so makes it more likely that we will form *individual* true beliefs. Theoretical rationality is also indispensable for placing these true beliefs in their broader informational context, for how we understand the world and how it relates to what we value. It is rarely the case that our decisions simply concern one particular belief in isolation; rather, in order to adequately understand our decision-making context, we often have to consider the extent to which a particular belief coheres with our other beliefs, about both descriptive and evaluative features of the world. Theoretical irrationality can undermine our understanding in this broader sense, even when it is compatible with the truth of a particular belief.[10] This suggests that delusions of the sort considered in the previous chapter can undermine decisional autonomy in two ways. They can either involve holding a false belief about an element of one's choice that is in fact decisionally necessary (in ways that I shall explain in later chapters) or delusional states can involve ongoing violations of norms of theoretical irrationality that otherwise jeopardize the individual's broader understanding.[11]

But theoretical rationality may also be said to enhance our *practical* autonomy as well as enabling our decisional autonomy. If our beliefs are true, the apparent reasons that ground our decisions are more likely to track our real reasons (rather than merely apparent reasons).[12] I am not here claiming that decisional autonomy requires that we *must* choose in accordance with our real reasons; this would make autonomy far too demanding for reasons explored above. However, when the apparent reasons that ground our decisions are more likely to reflect our real reasons, it is more likely that we will be successful in realizing the object of our desires.

To conclude this discussion of theoretical rationality and autonomy, I agree with Walker that a plausible minimal theoretical rationality condition of decisional autonomy may be phrased in the negative. We may plausibly say that decisional autonomy minimally requires the absence of theoretical irrationality, in so far as such

[10] For similar reasons, we may also be concerned about instances where doxastic justifications of true beliefs do not align with their propositional justifications. Indeed, this is why we should be concerned about what Shlomo Cohen has called the Gettier problem of informed consent. See Cohen, 'The Gettier Problem in Informed Consent'.

[11] Notice that this claim is quite compatible with the thought that delusions can be beneficial in some regards. Bortolotti et al. go further and argue that the fact that delusions do not undermine the capacity to form self-narratives suggests that delusions are compatible with self-governance (Bortolotti et al., 'Rationality and Self-Knowledge in Delusion and Confabulation'). However, whilst I agree that something like a self-narrative condition is a plausible condition of autonomy, it is not a sufficient condition. For reasons that I have discussed here, delusions can undermine decisional autonomy in ways other than undermining the capacity to form a self-narrative.

[12] Savulescu and Momeyer, 'Should Informed Consent Be Based on Rational Beliefs?'

irrationality is likely to lead us to (i) fail to hold a particular decisionally necessary true belief and/or (ii) to render us unable to place our beliefs in a broader coherent informational context that bespeaks understanding.

Of course, this claim will only be convincing if it is aligned with a criterion of understanding that sets out conditions on the decisionally necessary beliefs that individuals must hold in order to be able to make a particular autonomous decision. Yet, going beyond the minimum threshold condition of theoretical rationality, we may say that theoretical rationality can also make a positive contribution to decisional autonomy, in so far as it makes it more likely that we will hold crucial true beliefs. Further, in light of my claims in the preceding paragraph, I shall suggest in the next chapter that certain true beliefs may be necessary for the practical dimension of autonomy, that is, for us to be able to act effectively in pursuit of our ends. In any case though, contra Walker, the explanation for the role that theoretical rationality plays in autonomy goes beyond the fact that choosing on the basis of irrational beliefs involves choosing on the basis of an error.

2. Practical Rationality and Autonomy

A condition of theoretical rationality cannot explain why Jane the unwilling addict lacks autonomy. Jane can clearly be theoretically rational when she is acting on the basis of a compulsive desire. But is she being practically rational? And does this matter for her autonomy?

Rebecca Walker argues that the answer to both of these questions is 'yes'. Walker distinguishes between two kinds of goals that agents can have. Sometimes a goal is 'contingently' true of a person, in the sense that it is just a goal a particular individual has or chooses.[13] She gives the example of a person called 'James' who is a healthy weight, and who decides he wants to lose 10 pounds.[14] Crucially, she claims that there is nothing necessarily rational or irrational about such goals. In contrast, she contends that other goals may be rationally necessary (such as 'living well') or prohibited (such as 'self-destruction') for 'us as human beings'.[15]

In turn, Walker claims that these different kinds of goals are associated with different norms of practical rationality. With respect to rationally necessary goals, she suggests that one can be practically irrational simply by virtue of 'failing to recognise and choose in accordance with these goals'.[16] In contrast, with respect to contingent goals, practical rationality pertains only to the agent's willing the means that are necessary to achieve their goal. Failing to adhere to either of these norms can be sufficient for practical irrationality. In turn, since practical irrationality involves choosing on the basis of an error, practical irrationality undermines autonomous choice on Walker's approach.

Jane is most plausibly understood as acting irrationally in the first sense that Walker identifies. Recall that Jane wants to return to live a normal family life, and she knows that her drug-taking is jeopardizing this. If we understand her desire to return to a normal family life as a contingent goal, then her failure of practical

[13] Walker, 'Respect for Rational Autonomy', 342. [14] Ibid., 343. [15] Ibid. [16] Ibid.

rationality is a failure of doing what is necessary to achieve her desired end. Walker's theory also allows for the possibility that an agent who *endorses* their desire to take particularly dangerous drugs could also qualify as being practically irrational, if one holds that the avoidance of self-destruction is a rationally necessary goal.

Walker's account can thus offer us an explanation of why Jane is practically irrational. However, I believe that some misunderstandings in her conception of practical rationality lead her to overlook some other potential forms of practical irrationality, and to overplay others. These problems arise in part because Walker seems to conflate the distinction between objectivism and subjectivism about reasons, with the distinction between what I have called personal and impersonal reasons. More specifically, she adopts a subjectivist approach about reasons when discussing what she calls our contingent goals, and an overly narrow form of objectivism about reasons to what she calls our rationally necessary goals. As well as leading to an incomplete understanding of what errors of practical rationality might involve, we may also note that Walker's approach here is problematic for a deeper theoretical reason. It is not the case that objectivism or subjectivism about reasons are the sorts of theory that are true of some reasons but not others; rather these theories are about the fundamental grounding of *all* of our reasons.[17]

Consider first our 'contingent goals'. Walker takes the subjectivist line that such goals are not appropriate targets of rational assessment. However, it is entirely possible to offer an objectivist interpretation of these goals, and the norms of practical rationality that should apply to them. Indeed, 'contingent goals' bear a striking similarity to goals that an agent might have grounded by what I have called her 'personal reasons'.

Subjectivists and objectivists about reasons agree that practical rationality can demand that we should do the things that are necessary to realizing our desires. In accordance with a subjectivist view about reasons, this is essentially the *only* norm of practical rationality that Walker suggests is relevant for our contingent goals. However, objectivists can also not only offer a deeper justification for *why* this norm should obtain, they can also claim that our contingent goals themselves can be targets of rational assessment, given their relation to what I have called our telic reasons. Our desires to act in ways necessary to bring about some desired end can nonetheless lack rational justification, if the desire for the end in question is not itself rational.

To illustrate these different failings of practical rationality, return to Walker's example of James. Is it true that we can say nothing about the rationality of James' desire to lose 10 pounds, as Walker claims? Perhaps not; for instance, the objectivist might claim that James' goal is only rational if he decides to pursue this goal in response to his belief that he has some (perhaps self-interested) reasons to lose weight. Yet it is entirely possible that James does not adopt the goal as a rational response to such beliefs; it may just be a mere whim that he can't explain. Perhaps closer reflection would reveal to James that he does not actually care about losing the extra 10 pounds; he is after all already a healthy weight, and it will take an extreme

[17] Recall that this is the conclusion of Parfit's All or Nothing Argument.

amount of effort to lose the extra weight. In this permutation, James' contingent goal is one that he sustains in an *arational* sense.

It is also worth noting that contingent goals can be adopted *irrationally*, as an irrational response to reason-giving facts. Suppose for instance that Helen is on the brink of dying of starvation and yet still desires to lose 10 pounds—the objectivist might say that this contingent goal is irrational for Helen, even though it may not be irrational for James. The explanation for this is that there are facts that give Helen very strong self-interested reasons to avoid even limited weight loss, reasons which do not apply to James who *ex hypothesi* is a healthy weight.

The preceding discussion suggests that whilst an agent is practically irrational when failing to will the means necessary to achieving a contingent goal, she can also be practically irrational if she has adopted the contingent goal irrationally. In such cases, the agent will believe that she has strongly decisive reasons not to want the contingent goal. Alternatively, we may say that she may have adopted the goal arationally, on the basis of a brute desire that does not reflect what she actually cares about. This raises the question of whether autonomy is incompatible with practical arationality as well as practical irrationality. I shall defend the claim that it is below.

Walker's assumption of subjectivism about reasons with respect to our contingent goals leads her to overlook these potential deficits of practical rationality. I shall now suggest Walker's version of objectivism about reasons regarding our necessarily rational goals leads her to overplay apparent failures of practical rationality, and puts her theory in danger of collapsing into a substantive account of autonomy. According to Walker, a failure to choose in accordance with a necessarily rational goal is sufficient to qualify as a failure of practical rationality. Yet this is too strong. Objectivism about reasons is not committed to the claim that goals must be rationally necessary in this sense; we can have competing personal and impersonal reasons, and the truths about the relative strength of these reasons are highly imprecise. On more plausible versions of objectivism about reasons, one can be practically rational but fail to choose in accordance with a particular impersonal reason that we have 'as humans', as long as one is choosing in accordance with some *other* reason.

Julian Savulescu and Richard Momeyer make an even stronger claim about the apparent compatibility of autonomy and practical irrationality in their discussion of the following case:

> Assume that the harms of smoking outweigh the benefits. Jim has good reason to give up smoking. However, he may choose to smoke knowing all the good and bad effects of smoking.[18]

Savulescu and Momeyer use this example to illustrate their claim that 'a person may autonomously choose some course which he or she has no good reason to choose'.[19] In discussing this example, Savulescu and Momeyer claim that Jim's choice in this situation would be irrational; however, it would be autonomous if it were grounded

[18] Savulescu and Momeyer, 'Should Informed Consent Be Based on Rational Beliefs?', 283.
[19] Ibid.

by rational beliefs. On this reading then, their position seems to be that autonomous choice is compatible with errors of practical rationality.

Prima facie, many of Savulescu and Momeyer's claims in response to this example are appealing. Indeed, to claim that Jim *cannot* autonomously choose to smoke in this example might be understood to come close to endorsing a substantivist understanding of autonomy. Moreover, it also seems plausible to claim that Jim's choice to smoke would be practically irrational. Despite this, once we further unpack the example, I do not believe that it shows that autonomy is compatible with *all* forms of practical irrationality.

Let me take the points of agreement first; Savulescu and Momeyer equate 'having a good reason' with what I have called having a 'real reason'. Recall that such reasons do not depend upon the agent's beliefs (unlike their 'apparent' reasons). I wholly agree that autonomous choice is quite compatible with making errors about our *real* reasons; to claim otherwise would be to make autonomy all but impossible given the fact that we typically lack epistemic access to reason-giving facts that actually obtain.

However, this example does not establish that autonomy is compatible with *all* kinds of practical irrationality, or that practical rationality has no bearing on decisional autonomy. Much depends on how we flesh out the case. We are told that Jim knows all the good and bad effects of smoking. We are also told to assume that the harms of smoking outweigh the benefits, and that Jim thus has a real reason to give up smoking. Crucially though, we are *not* told whether Jim himself *agrees* with this impersonal ranking of the reasons associated with the harms of smoking and the reasons associated with its hedonic benefits for Jim. Yet, this feature is integral to understanding if we should understand his choice to be irrational in the manner that matters for autonomy.

The claim that Jim *is* autonomous with respect to his choice to continue smoking has more intuitive appeal when we assume that he does *not* agree with this impersonal ranking of values. In that way, his choice to smoke is a reflection of his own personal judgement about the relative strength of the reasons associated with the alternatives available to him; he values the pleasure smoking gives him over the longevity it threatens. It may be that his assessment of the strength of these reasons differs from the way in which others weigh them. Yet, further argument would be required to show that Jim would be evidencing a failure in practical rationality if he were to weigh his reasons in this way, particularly given the imprecise truths governing the strength of the reasons associated with different goods. This is not to say that such an argument would not be forthcoming. One way in which Jim may nonetheless evidence irrationality in weighing his values in this way is if he prioritizes the pleasure from smoking now over additional years of life in the future just because of an irrational bias towards what happens to oneself in the near future than in the more distant future.[20] Crucially though, an argument to that effect is necessary to establish the absence of decisional autonomy.

[20] For discussion about the nature of this bias, see Parfit, *Reasons and Persons*; Persson, *The Retreat of Reason*, ch. 15.

In contrast to those who conflate objectivism about reasons with acting in accordance with impersonal standards about impersonal reasons, the truths about the relative strength of many of our reasons, including those associated with the goods of health and pleasure, are imprecise. There is room for reasonable disagreement about which reasons are stronger. In so far as irrationality denotes a failure to act in accordance with clear and decisive reasons, it is hard to see how one could qualify as being irrational for simply holding a different view about which reasons win out in these cases.

To press the point further, suppose that someone like Jim, let's say Jimmy, *does* endorse the judgement in question; he agrees that he has stronger reasons to stop smoking than to continue, given what he knows about its harms and benefits. Yet Jimmy continues to smoke. It now becomes far less intuitively appealing to suppose that Jimmy is autonomous with respect to his decision to smoke; his action does not flow from his own evaluative judgements about what he ought to do.

In short then, whilst I agree with Savulescu and Momeyer that autonomy is compatible with errors concerning one's 'real' reasons, we should treat with caution their claim that Jim can be practically irrational and yet still be autonomous with respect to his decision. Much depends on whether we judge Jim to be irrational by some impersonal ranking of the strength of his reasons, or whether we judge Jim to be irrational given his own judgement about the strength of his reasons. Interestingly, my interpretation of the case seems broadly compatible with Savulescu's earlier work, which I delineated in my discussion of theoretical rationality above. In this earlier work, Savulescu claims that autonomy requires deciding on the basis of 'rational desires', that is desires that arise from an *evaluation* that the individual carries out in accordance with certain conditions (outlined above). The important point for my purposes here though is that it is the agent *herself* who must evaluate her options, in accordance with her own beliefs about the good.

This latter point raises an important feature of the rational autonomy view that I am outlining. Theoretical and practical rationality are not entirely separate domains of rationality. In particular, with respect to our thinking about autonomy, they are interlinked in the following important way: If we believe that decisional autonomy requires both theoretical rationality and practical rationality, in the sense that autonomous decision-makers must choose in accordance with what they believe they have sufficient reasons to do, then consistency demands that autonomous agents should be theoretically rational with respect to their evaluative beliefs about the strength of their different practical reasons. They must be receptive to reasons to think that some things have value, even if they do not need to prioritize a particular value in their decision-making. This feature will become important below, as I shall suggest that some agents who are practically rational may nonetheless lack autonomy because they are theoretically irrational with respect to their beliefs about what is good or valuable.

An objectivist account of reasons can thus offer a more nuanced account of how agents can be practically irrational. I now turn to the deeper question of why failures of practical rationality should be understood to undermine autonomy. As I have explained, for Walker the explanatory buck stops at the mere fact that practical

irrationality involves choosing on the basis of an error. I want to suggest that there are reasons for thinking that we should go deeper.

If one believes that errors of practical rationality undermine autonomy, and one also endorses objectivism about reasons (as I have assumed since Chapter 1), then one is committed to the claim that what matters for autonomy is acting in accordance with one's judgement about the relative strength of one's practical reasons. The deeper problem with simply saying that practical irrationality undermines autonomy because it involves choosing on the basis of an error, is that we still need an account of why we should trust that this aspect of our agency is the right place for the buck to stop with regards to autonomous decision-making. Why suppose that these judgements speak for us, or that they are the appropriate seat of self-government?[21]

To give a concrete example of the issue at stake here, Jillian Craigie has astutely observed that in some cases anorexic patients can express regret for their earlier refusals of treatment, and we may suspect that these patients' earlier decision-making suffered from a deeper kind practical irrationality. For some such patients, it is not the case that they were irrational because they failed to choose in accordance with what they valued (or what they desired); rather their regret for their earlier choices is grounded in the fact that they regret holding the values that undergirded their choices at the time of their decision, or for what they desired at that time.[22]

Craigie's example raises deep and important questions about the role of rationality in autonomy. I join her in believing that we can provide some of the answers to these questions by considering the positive role that practical rationality can play in a theory of decisional autonomy. It is not simply the case that practical irrationality undermines decisional autonomy because practical irrationality involves choosing on the basis of an error; rather, by fleshing out the objectivist approach to practical rationality that I have outlined here in certain ways, we can explain why our evaluative judgements about our rational desires can be the seat of autonomous decision-making. To make this argument, I will now consider how rationalist theories of autonomy have been developed in the wider philosophical literature concerning the philosophy of action. This, and my discussion of controlling influences in Chapter 3, will provide me with the necessary platform to engage with Craigie's discussion in more detail when I turn to the issue of rational competence in Chapter 8.

3. Values, Identification, and Authority

Questions concerning which aspect of our agency constitutes the source of our autonomy have been widely discussed in philosophy of action. In this section, I shall briefly trace the history of this discussion before defending the account I favour in the following section. Readers who already are familiar with the philosophical literature on autonomy and identification may wish to skip this section.

[21] Note that a similar problem will arise for subjectivist but with respect to why one's desires should play this role.

[22] Craigie, 'Competence, Practical Rationality and What a Patient Values'.

In an influential paper that has somewhat set the terms of the debate in this area, Harry Frankfurt sought to answer the question of why agents who act on compulsive desires (like Jane the unwilling addict) seem to lack freedom of the will.[23] The explanation for this, on Frankfurt's account, is that such agents do not *identify* with their motivating desires.

According to Frankfurt, conscious entities have 'first-order desires' to do or have certain things. Some of these desires are the ones that actually motivate them to act; for Frankfurt, it is these 'effective' first-order desires that constitute 'the will'.[24] So, on this view, if I have a desire to *x* and I end up *x*-ing, this particular first-order desire is effective, and thereby constitutes my will. In so far as we are creatures that have such first-order desires, nothing separates humans from other members of the animal kingdom. However, according to Frankfurt, 'persons' are unique in that they can also have 'second-order desires'; these desires are 'higher order' desires that have as their object a certain first-order desire.[25] Further, persons can have second-order *volitions*; such volitions are a particular species of second-order desire, defined by their object. The object of these volitions is that a particular first-order desire becomes *effective* in moving them to act.[26]

The relationship between the agent's effective first-order desires, and their second-order volitions is integral to freedom of the will for Frankfurt. He writes:

> . . . it is in securing the conformity of his will to his second-order volitions . . . that a person exercises freedom of the will.[27]

Frankfurt's approach can seemingly explain why being alienated from one's motivating desire can undermine autonomy. Recall the example of Jane from the introduction.[28] We can understand Jane as having two conflicting first-order motivating desires; she has an urge to take drugs, but she also harbours a desire not to do this. We can also understand her as having a second-order volition: for the latter first-order desire (to refrain from this behaviour) to constitute her will. Nonetheless, Jane's first-order desire to take drugs becomes effective; accordingly, she lacks autonomy with respect to her drug-taking on Frankfurt's approach. We may contrast Jane with another addict Beatrice who would be autonomous on Frankfurt's approach: suppose that Beatrice has only one first-order desire; she wants to take drugs. However, unlike Jane, suppose that Beatrice's second-order volition is that this desire *should* come to constitute her will. She embraces her addiction, and would reinstate her first-order desire to engage in this behaviour should it wane.[29]

The reason that Beatrice is autonomous with respect to her drug-taking on Frankfurt's approach is because her motivating desire is authentic to her in a way that Jane's is not. At least by the lights of his original theory, Beatrice's identification

[23] Frankfurt's intention in the work was to provide a theory of freedom of the will and its relation to personhood, rather than autonomy per se. However, this has not prevented many commentators regarding his theory as a prominent example of a theory of autonomy. Taylor is a notable exception. See his arguments against this interpretation in Taylor, *Practical Autonomy and Bioethics*, ch. 3.

[24] Frankfurt, 'Freedom of the Will and the Concept of a Person', 8. [25] Ibid., 10. [26] Ibid.

[27] Ibid., 15. [28] Introduction, 00.

[29] These examples correspond to Frankfurt's examples of the unwilling addict and the willing addict. Frankfurt, 'Freedom of the Will and the Concept of a Person', 12–15.

with her desire ensures that it is a reflection of what she really wants, or of the central elements of her 'true self'.

Given the controversial nature of 'the self',[30] it is perhaps apposite here to clarify the role that the concept is playing here.[31] On this understanding it is not merely a 'grammatical error'[32] to claim that agents have a self in some sense; rather the self can be understood as the metaphorical locus of the agent's 'character',[33] or of the psychological continuities that ground personal identity on some theories.[34] In holding that the self is something that both persists over time and can undergird the intelligibility of the agent's long-term diachronic plans, this understanding of the self is naturally not compatible with those theories that deny that the self can persist over long periods of time,[35] or in a diachronically continuous sense.[36] However, it is compatible with a number of claims that are incorporated into a diverse range of theories of the self. Most critically, it is not committed to the contentious claim that the self is static, or an extant metaphysical essence;[37] the true self can be construed to persist even if the elements that constitute it change over time, as long as the agent changes them in accordance with the sorts of procedure that procedural theories of autonomy seek to explicate.[38]

Frankfurt's theory has been highly influential, and it is still appealed to in bioethical discussions of autonomy. However, it faces a similar question to the one that I raised about rationalist theories of autonomy in bioethics at the end of the previous section. Why should we trust that our second-order volitions should serve as the proxy for the 'true self' and as the seat of self-governance? Here, it seems Frankfurt faces a choice between two unappealing alternatives. First, perhaps an *even higher* order volition authenticates one's second-order volitions as being one's own. However, this reply is problematic because it seems to lead inexorably to a regress

[30] For instance, Ekstrom writes '... in order to understand autonomous action ... we need a working conception of what constitutes the "self"' (Ekstrom, 'A Coherence Theory of Autonomy', 599). In contrast, Berofsky argues against conceiving of autonomous agency as that which proceeds from some extant metaphysical self. See Berofsky, *Liberation from Self*.

[31] For a deeper discussion of the role of the self in conceptions of authenticity, see Friedman, *Autonomy, Gender, Politics*, 3–29.

[32] Kenny, *The Self*, 4.

[33] Both Mill and Aristotle invoke the agent's character as a ground of choice in their discussions of individuality and voluntariness respectively. See Mill, *On Liberty*; Aristotle, *Nicomachean Ethics*, book III. See Meyer, 'Aristotle on the Voluntary' for a useful discussion of how character relates to voluntariness in Aristotle's theory of virtue.

[34] See Parfit, *Reasons and Persons* (Part Three) for a classic psychological theory of personal identity. Michael Bratman explicitly points out that the self-governing policies that undergird autonomy on his view are inextricably related to the agent's identity, since they concern plans that are constituted by psychological continuities. See Bratman, 'Planning Agency, Autonomous Agency', 41.

[35] For example, see Strawson's 'Pearl View of the Self' in Strawson, 'The Self', 424. For an explanation of how Strawson's and David Hume's seminal view differ, see Strawson, 'Hume on Himself'.

[36] For example, see Hume's exposition of his so-called 'Bundle Theory of the Self'; Hume, *A Treatise of Human Nature* (section entitled 'Of Personal Identity'). For a rejection of the Strawsonian and Humean approaches to the self, see Olson, 'There Is No Problem of the Self'.

[37] For criticism of this essentialist view, see DeGrazia, *Human Identity and Bioethics*, 233–4.

[38] For further discussion of this understanding of the self and how it functions in the kind of account of autonomy that I develop here, see Pugh, Maslen, and Savulescu, 'Deep Brain Stimulation, Authenticity and Value'.

of increasingly higher order conative attitudes. Alternatively, he might claim that at some level, a higher order desire cannot be authenticated, and does not require authentication.[39] However, this reply leads to what John Christman terms the *ab initio* problem,[40] since it implies that the authenticity of one's first-order desires can only be ensured by a second- (or higher) order desire that is not itself authentically the agent's. As Christman puts it, this would involve the claim that '. . . desires can be autonomous without foundations',[41] and this, he claims, renders the second response 'implausible'.[42]

One might suppose that the problem with Frankfurt's theory here is its over-reliance on non-cognitive elements as constituting the true self. One reason for doubting that our second-order volitions in particular can constitute the true self is that agents can, as Frankfurt concedes, form these desires in a capricious manner, and without any serious consideration.[43] If these volitions are thus 'blind or irrational'[44] impulses, then it is hardly surprising that they cannot serve as an appropriate seat of self-governance. In contrast, one might suppose that a rationalist theory would not fall foul of the same problem because reason allows us to identify the good in our evaluative judgements, and our rationally warranted desires are thus not blind impulses.[45]

Yet such appeals to the authority of 'rationality' will not be sufficient unless one discounts the possibility that agents could similarly be alienated from their values; why suppose that our values constitute the real self? This objection has more and less plausible variants. First, one might object that the prospect of alienation can arise because rationalist theories entail that the agent's values must track some objective good, and that they are thus unable to account for the undeniable fact that we '. . . sometimes place value on senseless or masochistic ends, that is, ends that have no objective value'.[46] However, my discussion of objectivism about reasons, real and apparent reasons, and the difference between personal and impersonal reasons should make it clear that a suitably nuanced theory of rationalist autonomy need not fall foul of this form of the objection.

However, the objection can be raised in a more nuanced and fundamental way. David Velleman writes:

> The agent's role cannot be played by any mental states or events whose behavioural influence might come up for review in practical thought at any level. And the reason why it cannot be played by anything that might undergo the process of critical review is precisely that it must be played by whatever directs that process. The agent, in his capacity as agent, is that party who

[39] This is the horn of the dilemma that Frankfurt grabbed in his later work, appealing to the concepts of decisiveness and satisfaction. See Frankfurt, *The Importance of What We Care About*, 21; Frankfurt, *Necessity, Volition, and Love*, 104.

[40] Christman, 'Autonomy and Personal History', 7. [41] Ibid. [42] Ibid.

[43] Frankfurt, 'Freedom of the Will and the Concept of a Person', 13, note 6.

[44] Watson, 'Free Agency', 208.

[45] For an early rationalist response to Frankfurt in this vein, see Watson, 'Free Agency'.

[46] Berofsky, *Liberation from Self*, 80. Berofsky's complaint here is most readily raised against rationalist theories that employ a Platonic conception of objective goods. See Watson, 'Free Agency'; Wolf, *Freedom within Reason*.

is always behind, and never in front of, the lens of critical reflection, no matter where in the hierarchy of motives it turns.[47]

In light of these remarks, Velleman posits that agency is not grounded in some collection of psychological elements that constitute a 'true self'; rather, it must be grounded by a motive that is never subject to critical reflection, and which is nonetheless functionally equivalent to the agent herself. Initially, he identifies this motive as the fundamental concern that all agents share to act in accordance with reasons, where reasons are 'considerations by which an action can be explained and in light of which it would therefore make sense to the agent'.[48] In more recent work, Velleman has further specified his understanding of the constitutive motive of agency in accordance with this understanding of reasons. On the further developed view, the constitutive inclination of agency is not merely the inclination to act for reasons, but rather the inclination to acquire self-understanding, that is, the inclination to render oneself 'intelligible' in the folk psychological sense.[49] I shall use the latter understanding in my discussion below.

Whilst Velleman offers an account for how the rationalist might respond to the problem of alienation, I shall not pursue it further here for two reasons. First, aspects of the view are in tension with objectivism about reasons that grounds the theory of rational autonomy that I am developing here. On Velleman's understanding, our reasons for action only apply if we have the higher order inclination to render our actions intelligible. In making this claim, he is seeking to forge a middle ground between objectivism and subjectivism about reasons, insofar as our reasons still depend on a particular inclination, but one that is central to understanding ourselves as agents. However, the subjectivist element of this claim still seems open to Parfit's criticism of such theories; in particular, one might object that one's reason to avoid a period of agony is not merely contingent on whether one has the inclination to render one's actions intelligible. Moreover, Velleman's articulation of the nature of our reasons for actions also incorporates subjectivist commitments. On Velleman's view, holding a particular lower order desire for some outcome, in conjunction with a higher order desire for my actions to be intelligible, is sufficient for having a reason to act to bring about this outcome; the two desires can explain the action in the required sense. However, for reasons I explored in Chapter 1, the objectivist will find this claim and the absence of evaluation in this model problematic; for the objectivist, it is crucial that our practical reasons *justify* our actions, rather than merely explain them.

However, the second more fundamental issue with this approach is that we may doubt the underlying premise that motivates it, namely that the only mental state that can have the authority to speak for the agent is one that is itself not subject to critical review. This assumption motivates Velleman's claim that agency requires an inclination towards self-understanding. This is not only a considerable theoretical commitment about the nature of agency, it is also empirically dubious that the individuals that we would typically categorize as agents all share this inclination. Yet, rather than making this assumption, one might instead claim that the relevant psychological

[47] Velleman, 'What Happens When Someone Acts?', 477.
[48] Ibid.
[49] Velleman, *How We Get Along*.

elements can plausibly receive agential grounding, even if they themselves can be subject to critical review. Instead of being grounded by some fundamental and 'pure' inclination as Velleman claims, it seems plausible that certain psychological elements could be mutually reinforcing and justificatory. Not only that, but these mutually reinforcing psychological elements can plausibly have agential authority just because they constitute our practical identities.[50] This is the thrust of Laura Waddell Ekstrom's coherence approach, a modified version of which I shall defend in the remainder of this chapter.[51]

4. Defending a Modified Coherence Approach to Rationalist Authenticity

On the coherence approach, an agent is autonomous when they act on a first-order desire if they have a 'personally authorized preference' for that desire to be effective. This terminology requires some explanation. First, a 'preference' in this context is understood to be a desire for a certain first-order level desire to be effective in moving the agent to act. However, this understanding of a preference moves away from a Frankfurtian picture of second-order volitions, since a preference on this account is formed in accordance with the agent's subjective conception of the good. Crucially, although this evaluation need not occur at the conscious level, it must actually be performed at some point.[52] Notice then that this understanding of 'preferences' is compatible with objectivism about reasons; the point is the agent forms her preferences on the basis of her beliefs about what is valuable. Such preferences are thus grounded by reasons of the sort that objectivists champion and subjectivists deny.

Second, a preference is *personally authorized* if it 'coheres' with the agent's 'character system',[53] that is, the agent's set of preferences at time *t*, in conjunction

[50] Michael Bratman adopts a similar approach in Bratman, 'Planning Agency, Autonomous Agency'. However, like Velleman, Bratman's approach incorporates a number of important subjectivist assumptions. In particular, he is quite clear that planning attitudes that undergird autonomous agency do not need to be grounded by an evaluative judgement. For Bratman's own understanding of valuing, see Bratman, 'Valuing and The Will' and Bratman, 'Identification, Decision, and Treating as a Reason'.

[51] Although I am choosing to explicate autonomy in terms of authenticity conditions, an alternative approach from the moral responsibility literature that is amenable to my rationalist approach claims that moral responsibility requires some form of reasons responsiveness. See Fischer and Ravizza, *Responsibility and Control*; Haji, *Moral Appraisability*. In turn, reasons responsiveness requires that agents are both receptive and reactive to a broad set of reasons. An agent is *receptive* to reasons if they are able to identify and process good reasons. An agent is *reactive* to reasons if their decision-making mechanism would give rise to different action in some hypothetical cases where different reasons obtained. Although I am sympathetic to the claim that autonomy requires reasons receptivity, I am less certain that it requires reasons reactivity. For discussion, see Mele, 'Fischer and Ravizza on Moral Responsibility', 288–94. For other objections to the claim that autonomy (as opposed to responsibility) requires reasons reactivity, see Christman, *The Politics of Persons*, 141.

[52] Ekstrom, 'A Coherence Theory of Autonomy', 603.The criterion of unconscious reflection is presumably a pre-emptive defence against the charge that people do not typically reflect on their motivational states at the time of action. In a similar vein, Savulescu appeals to the claim that autonomy is a dispositional property—evaluative reflection must occur at some point, although it need not be at the point of action. See Savulescu, 'Rational Desires and the Limitation of Life Sustaining Treatment', 199–200.

[53] Ekstrom, 'A Coherence Theory of Autonomy', 606.

with the set of propositions that the agent accepts at *t*. The latter are termed the agent's 'acceptances', and are beliefs formed in accordance with the individual's subjective conception of the true.[54] Finally, a preference for a particular desire to be effective coheres with an agent's character system if it is either (i) more valuable for the agent to prefer that desire than it is for her to prefer a competing desire, on the basis of their character system, or (ii) as valuable for the agent to prefer the conjunction of that desire and another neutralizing desire *n*, as it is for her to prefer a competing desire.[55]

It is important to clarify an ambiguity here regarding Ekstrom's terminology of it being 'more valuable for an agent to prefer a *desire*'. It may be more valuable for an agent to prefer one thing to another for two different kinds of reason. The reason might be object-given in the sense I have been so far considering; that is, the object of one desire can be more valuable than another. Alternatively, it could be more valuable for an agent to prefer a desire because she has *state-given reasons* to hold a particular preference. To illustrate, suppose that someone threatened to torture you unless you held a particular desire. This would give you a state-given reason to be in the state of holding the desire in question, even if you had no object-given reason to want the object of the desire itself (suppose that you would be tortured unless you held a desire to do something you find repulsive). Whilst Ekstrom's choice of terminology might lead one to think that she is appealing to state-given reasons, I think it is most natural to understand her view as appealing to object-given reasons.[56]

There is much to be said in favour of the coherence theory. It can explain why those of an agent's preferences that cohere with her character system have agential authority, in so far as the agent's coherent preferences and acceptances may plausibly be understood as representing the agent's 'true self'. There is a strong case in favour of this view, since cohering elements of the self are likely to be 'particularly long lasting',[57] since they are 'well-supported with reasons'.[58] By virtue of this support, they will also be 'fully defensible against external challenges',[59] as well as being preferences that the agent feels 'comfortable owning'.[60] However, our characters are not thereby static; elements of our character systems can and do change. Crucially though, if new elements of our psychological economies are to cohere, then they must admit of rational justification in accordance with other elements of our characters. Accordingly, character change that is compatible with autonomy will be gradual, and akin to rebuilding Neurath's raft. New elements have to fit with the pre-existing structure; moreover, replacing the complete structure wholesale in one fell swoop would rupture the continuity of the agent's identity.

The coherence approach can thus offer a model for how our evaluative judgements about what we have reasons to do can be understood as the appropriate seat of self-governance. Practical rationality, understood as acting in accordance with our

[54] Ibid. [55] Ibid., 611.

[56] See Parfit, *On What Matters*, 50–2 and appendix A for further discussion of state-given reasons. Notably, he is sceptical about the import of such reasons, if they do in fact obtain in separation from object-given reasons.

[57] Ekstrom, 'A Coherence Theory of Autonomy', 608. [58] Ibid. [59] Ibid. [60] Ibid., 609.

rationally warranted preferences, has a positive role in autonomy because our preferences are abiding elements of our characters, and thus have agential authority. To conclude this chapter, I shall defend the coherence approach from three objections, and in doing so slightly refine the view. To be clear, in light of the distinction I drew in the introduction, the objections I consider here are objections to the claim that a rationalist coherentist condition should feature amongst the conditions of what *constitutes* decisional autonomy. I shall consider further objections to the implications of my theory for the *causal* conditions of autonomy (including objections grounded in concerns about demandingness and the role of emotions in rationalist autonomy), in Chapter 7.

(i) An Asymmetry of Theoretical and Practical Rationality?

On Ekstrom's description, the coherence theory allows for a possible asymmetry between theoretical and practical rationality. Preferences have to be rationally warranted in the objectivist sense that the agent must believe that they have reasons to have those preferences, reasons that are based upon their subjective beliefs concerning the good. Yet, acceptances need only be held in accordance with the agent's 'subjective conception of the true'; as such, the coherence theory denies that the autonomous agent's beliefs must be in any way rationally warranted. However, this asymmetry between practical and theoretical rationality is problematic. In section 1, I argued that autonomous decision-making plausibly requires a degree of theoretical rationality. Towards the end of section 2, I also mentioned that if one accepts this point, then consistency demands that we should claim that autonomous agents should be theoretically rational with respect to the evaluative beliefs about the good. Crucially, these evaluative beliefs significantly ground our practical rationality. This point comes to fore in cases where agents might plausibly lack autonomy with regards to their motivating desire, not because the motivating desire itself is incongruous with their subjective conception of the good, but rather because the agent's *beliefs* about the good are theoretically irrational. For instance, as Fulford explains, delusions that threaten autonomy can be evaluative and not simply factual.[61]

To illustrate this point, consider a sufferer of clinical depression. In some cases of this psychiatric disorder, the sufferer may have a suicidal desire that they personally authorize; but this authorization may stem from a belief about the disvalue of their own life that they hold irrationally.[62] In saying that the belief is irrational in this sense, I do not mean to say anything about whether the content of the belief is objectively true or false. As I observed in the previous chapter, it is a mistake to assume that delusions are necessarily false; so accepting that evaluative delusions are possible does not entail the claim that they involve *false* evaluative judgements. Moreover, as I shall clarify in Chapter 8, I am not denying that a desire to end one's own life can never be practically rational, nor grounded by theoretically rational evaluative beliefs. The point here is rather that in some cases agents are not themselves able to offer any cogent reasons as to *why* they hold certain dubious evaluative

[61] Fulford, 'Evaluative Delusions'. [62] See Beck, *Depression*, 3.

beliefs (perhaps about their own self-worth), or to respond to any epistemic reasons with which they are presented against holding such beliefs (for instance, evidence that other people care about them, contrary to their own impression). Instead, they may simply adopt these beliefs unshakeably in a manner that bespeaks a delusional state.

There are important questions about how we should delimit the scope of the concept of an 'evaluative delusion' to which I shall return in Chapter 8. Moreover, we should of course take care not to automatically assume that sufferers of psychiatric disorder always lack autonomy with respect to desires that constitute a significant diagnostic criterion of their condition (again, a point to which I shall return). Yet, it seems plausible to claim that the agent in the particular case under consideration plausibly does lack autonomy with respect to their suicidal desire given the nature of the evaluative belief upon which it is based; I suggest that just as we believe that an agent can lack autonomy if they are compelled by a motivating desire from which they feel alienated, so too can an agent lack autonomy if their endorsement of their motivating desire is based upon an irrational belief about the good. In cases such as the one I am considering here, it seems possible that an agent's decisional autonomy can be undermined by delusional evaluative beliefs, as well as compulsive first-order desires.[63]

Whilst the coherence theory's appeal to a purely subjective understanding of the truth with respect to the agent's acceptances is problematic for this reason, the problem is easily remedied. Like rationalist theories of autonomy in bioethics, the coherence approach should adopt a condition of theoretical rationality with respect to acceptances.

(ii) Competing Desires and Coherence

The second objection pertains to the criteria of what it is for an agent's preferences to cohere. On Ekstrom's view, when agents have to decide which of two competing preferences should win out and cohere with their other central preferences, the autonomous agent will decide that one preference defeats another on the basis that it is *more valuable* for her to prefer the object of desire *d* to the object of desire *g*, or at least as valuable to prefer the object of desire *d* and some neutralizing desire *n*.

A problem with this view is that it is unable to account for the possibility that an agent could be autonomous with respect to a desire to act in manner that they believe to be sub-optimal. Consider the following example. Suppose that Jim values having a career in medicine, but also values spending more time with his family. Following a great deal of consideration, let us suppose that Jim forms the judgement that it would be slightly more valuable for him to prefer that his desire to spend more time with his

[63] Radoilska argues that depression can undermine decisional autonomy because it involves paradoxical identification, in which one identifies with what one loathes, in this case, oneself. See Radoilska, 'Depression, Decisional Capacity, and Personal Autonomy'. I am sympathetic to this view, and acknowledge that this is a related way in which depression can serve to undermine decisional autonomy.

family be effective in moving him to act.[64] Would it really be the case that, having made this assessment about what is more valuable (and sticking to it), Jim would no longer be autonomous with respect to his decision if he became motivated to instead pursue a career in medicine? Admittedly, he would be doing so in the knowledge that he could be doing something else that he believed to be slightly more valuable; however, it still seems plausible to claim that Jim could nonetheless still be autonomous with respect to this decision.[65] Notice that this is compatible with the claim that Jim would have been *more* autonomous if he had chosen to act in accordance with what he believed to be his strongest reasons. The point that I am making here is that it is more plausible to make these two claims, rather than to rule out the possibility of Jim's autonomy here.

Indeed, as I have stressed throughout this chapter, truths regarding the relative strength of our different competing self-interested reasons can be highly imprecise, and there may just be no clear way of deciding which of two competing preferences *A* and *B* it would be more valuable for one to prefer. Paul Hughes has argued that when a person acts from volitional ambivalence like this:

> ... she is not autonomous either with respect to the desire that prompts her action or the action itself ... [since] ... in cases of volitional ambivalence there is no single conative 'self' directing the agent's actions.[66]

Hughes seems to be making a similar assumption to Velleman here in appealing to the need for a single conative self directing critical evaluation. As I have explained, the coherence approach can explain why this is not necessary; cohering elements of our character systems admit of both rational justification, and mutually reinforcing justifications. Moreover, *pace* Hughes, it is not clear why an agent in this situation would not be autonomous with respect to their action once they had elected to act in accordance with, say, preference A rather than preference B. Once the agent has plumped for A, it seems plausible to claim that they will be autonomous with respect to acting in pursuit of A *in so far as preference A is itself still rationally warranted*. In plumping this way, otherwise ambivalent agents simply act in a manner that serves to constitute their will.[67] Although A is no better or worse than B, this only means that the agent may lack a rational basis for their choice of A over B; but this does not mean that they lack autonomy with respect to their acting in pursuit of A, since that act itself is still rationally warranted.[68] The choice of A over B is thus a choice to prioritize a certain set of reasons over another, and to emphasize the corresponding

[64] For simplicity, I am assuming here that the beliefs that Jim knows he has at the time of deliberation prior to this judgement exhaust all of the beliefs he has relevant to this decision. However, as Arpaly points out, this need not be the case. See Arpaly, 'On Acting Rationally against One's Best Judgment'.

[65] Sher discusses a similar example. Sher, 'Liberal Neutrality and the Value of Autonomy', 143. See also Raz, *The Morality of Freedom*, 304.

[66] Hughes, 'Ambivalence, Autonomy, and Organ Sales', 238–9. Bratman also raises this sort of underdetermination case as raising a concern for theories of autonomy that appeal to rationalist considerations. See Bratman, 'Reflection, Planning, and Temporally Extended Agency'.

[67] Ruth Chang defends a similar view, and a detailed account of the nature of what she calls 'hard choices' in Chang, 'Hard Choices'.

[68] Sher, 'Liberal Neutrality and the Value of Autonomy', 144 makes a similar point.

aspects of one's character system, even whilst acknowledging that the alternative choice also represents elements of one's character system that one would not otherwise repudiate.[69] An agent's chosen preference can still cohere with her other central preferences and acceptances, and ground an autonomous choice in such circumstances. Indeed, in light of this discussion, it is notable that in medical contexts, a patient's demonstrable ambivalence in the face of difficult choices in an end of life decision-making context is not understood to readily undermine decision-making competence.[70]

In light of the above, the coherence approach should adopt the weaker claim that autonomous agents should choose in accordance with preferences that they have a sufficient reason to adopt. On this understanding, coherence is incompatible with irrationality in Parfit's sense, but not with choosing in a less than fully rational manner. With this amendment, the coherence approach can accommodate the plausible claim that autonomous agents can make sub-optimal choices, which may still reflect central elements of the agent's characters, particularly in the light of the imprecise truths governing the strength of our competing practical reasons.

(iii) Authentic Alienation?

Suzy Killmister has recently raised an important challenge for rationalist theories of autonomy that is apposite here.[71] She asks us to consider a case in which an agent accepts that a motivational attitude they hold is irrational, but which they nonetheless regard as providing them with sufficient justificatory reasons to act, because it reflects what they take to be their true self. Most theories of autonomy, she claims, cannot account for the thought that such an agent seems to be autonomous along some dimensions, but less autonomous along others.

To give a concrete bioethical example, an anorexic patient might regard her desire to refrain from eating as irrational, and yet also regard it as providing her with sufficient justificatory reason to refrain from eating. The justificatory reason arises from the fact that the patient understands this irrational desire to partly constitute her real self.[72] Killmister claims that in order to account for this sort of case, we need to split what I am calling the reflective element of autonomy into two components, which she terms 'self-definition' and 'self-realization'. Self-definition pertains to the reasonableness of an agent's attitudes, whilst self-realization pertains to the extent to which the agent's intentions track what she takes herself to have most reasons to do.

However, the rationalist theory developed here can also provide a theoretical basis for those who are ambivalent with regards to such an agent's autonomy, providing certain assumptions are met. The theory can also provide a basis for critiquing the intuition that the autonomy of such an agent is in some way 'mixed'. To see why, we need to think more deeply about both the nature of the reasons and the conceptions of rationality in play in Killmister's example. Recall that the patient in this example

[69] Joesph Raz argues that such choices play a particularly important role in shaping our character. See Raz, *Engaging Reason*, 242.

[70] Gavaghan, 'In Word, or Sigh, or Tear', 248.

[71] Killmister, 'The Woody Allen Puzzle'.

[72] For some empirical support for the plausibility of such an example, see Tan et al., 'Competence to Make Treatment Decisions in Anorexia Nervosa'.

regards her desire to refrain from eating as irrational. Nonetheless, she takes that desire as providing her with reasons for action, in so far as she regards that desire as partly constitutive of her authentic self. On Killmister's model, such a patient would lack autonomy in one sense, that is, with respect to her self-definition (in so far as her authentic self incorporates elements that she herself takes to be irrational). Yet she would also, in some sense, be autonomous with respect to her self-realization, in so far as she is acting in accordance with what she believes she has most reason to do, that is, act in accordance with her authentic self.

On the framework that I have presented here, the plausibility of such a patient being 'mixed' with respect to their autonomy in this way relies on two assumptions. First, that an agent's authentic self could incorporate attitudes that she herself takes to be irrational. Second, that acting in accordance with one's authentic self *for its own* sake can be regarded as good in a reason-implying sense. I shall consider each in turn.

Our views regarding the plausibility of these assumptions are likely to be complicated by different interpretations of the 'true self'. On some understandings of authenticity that are implicit in philosophical theories of autonomy, the true self is understood as being perpetually created; living authentically is a matter of consciously shaping one's own character in accordance with one's desires and values.[73] In contrast though, one might endorse an alternative essentialist understanding of authenticity, according to which the true self is an extant and largely static essence that we need to discover rather than create.[74]

The claim that the above anorexic patient is partly autonomous in the way that Killmister understands her to be seems to rely on an essentialist conception of authenticity.[75] The reason for this is that many existentialist understandings of authenticity would most likely reject the first assumption outlined above: If, as existentialist approaches maintain, it is the agent herself who decides how to shape her authentic self on the basis of her own values and conception of the good, then it is not clear how the authentic self could be understood to incorporate elements that the agent *herself* takes to be irrational. Notice though that this claim is compatible with the thought that an existentialist conception of the true self might plausibly incorporate elements that one believes others will deem to be irrational. Indeed, some anorexic patients might claim that their true selves incorporate irrational elements in this normative sense, in so far as they might admit that it would be more rational to prioritize their health over a low weight, from a third-party perspective. Crucially though, this need not commit such patients to regarding their desire to maintain a low weight as irrational; they may yet believe that they are acting in accordance with what *they* have strongest reason to do.

[73] For a discussion of the existentialist approach and autonomy, see DeGrazia, *Human Identity and Bioethics*, ch. 3.

[74] For discussions of this distinction, see Levy, 'Enhancing Authenticity'; Pugh, Maslen, and Savulescu, 'Deep Brain Stimulation, Authenticity and Value'.

[75] Interestingly, Killmister's interpretation of the anorexic case here runs contrary to an essentialist tradition in bioethics that claims that the 'anorexic self' must be *inauthentic*, on the basis that it is grounded by pathological or self-destructive values. See Tan et al., 'Competence to Make Treatment Decisions in Anorexia Nervosa'; Nordenfelt, *Rationality and Compulsion*. I shall discuss such accounts in greater detail in Chapter 8.

The rationalist theory that I have developed here can accommodate the thought that such individuals can be autonomous with respect to desires that are rationally endorsed in this sense, as long as the beliefs about the good upon which they are based are not held in a theoretically irrational manner. Yet, if this is the correct interpretation of Killmister's view, then the agent is not mixed with regard to her autonomy; she is acting in accordance with a desire that she rationally endorses, even though she acknowledges its apparent irrationality. Yet this just means that she disagrees about the strength that we ought to attribute to different reasons; this alone is not sufficient for practical irrationality as I have described it in this chapter.

The authentic self cannot incorporate irrational elements in this motivational sense on a number of plausible existentialist understandings of authenticity. However, this is not a problem for the essentialist understanding; one's static essence may incorporate attitudes that one now takes to be irrational on the basis of one's own beliefs about value, or one's beliefs about what one *should* value according to impersonal criteria. So Killmister's suggestion that the agent in her case partly lacks autonomy seems to rely on an implicit essentialist conception of authenticity.

The second question is whether this essentialist conception of authenticity can provide a sufficient justificatory reason for action, as Killmister's second assumption requires. Naturally, the first potential problem with this claim is that the essentialist conception of authenticity is somewhat contentious, in so far as it seems to rely on the assumption that we have a deep, immutable, hidden essence that is immune to our own evaluative stance.[76] Notwithstanding this issue, the essentialist understanding also owes us an account of why living authentically on this conception should be regarded as good in a reason-implying sense. Even assuming that an essential self exists, to claim that this essence must be good and that it ought to be promoted without further argument seems to come close to making the naturalistic fallacy.

Suppose, though, that such an account can be provided;[77] if living in accordance with an essentialist conception of the self can be construed as good in a reason-implying sense, and that conception of the self incorporates elements that the agent herself takes to be irrational, then it seems that Killmister's ambivalent intuition about the agent's autonomy in this case can be compatible with the rationalist account that I have developed here. However, we may notice that the strength of one's reason to live in accordance with this essence on such an account would also have to be particularly strong. After all, the reasons associated with living authentically would need to be sufficient to compete with the agent's reasons to pursue other goods, perhaps even including survival in the case of severe anorexia.

The rationalist framework I have outlined here can not only account for Killmister's own ambivalent intuition in such cases, but it can also account for the possibility that we may not find the intuition about ambivalence compelling. Whether we share Killmister's intuition will depend on our credence in essentialist conceptions of authenticity, their value, and the possibility that an agent could

[76] Strohminger, Knobe, and Newman, 'The True Self'; DeGrazia, *Human Identity and Bioethics*, 233–4.

[77] For a classic defence of essentialist authenticity as a normative ideal in this respect, see Taylor, 'The Ethics of Authenticity'. For some considerations that speak in favour of this approach in the context of mental disorders, see Erler and Hope, 'Mental Disorder and the Concept of Authenticity'.

rationally prioritize this value over other goods that may be in play in her decision-making.

In contrast to the essentialist conception of authenticity upon which Killmister's case seems to rely, the coherence approach I have outlined in this chapter draws on both essentialist and existentialist themes in the conception of authenticity that it invokes. From the essentialist tradition, it takes the claim that we may have certain more or less fixed elements that partly constitute our character system. From the existentialist tradition, it takes the claim that we may be able to choose which of these more or less fixed elements to bring to the fore in a coherent nexus, and which to downplay on the basis of the web of values that we come to develop. On this approach, if the individual herself believes that a certain element of her character system is not valuable, as Killmister's anorexic patient does, then this element of her character system is inauthentic, and cannot be understood as a suitable ground of autonomous decision-making. However, as I shall explore in Chapter 8, the theory is also compatible with the thought that some anorexic patients may coherently experience their disorder as a part of their authentic self. In this part of the book, I shall also return to Craigie's concern that such patients may also later regret the values that previously informed their decision-making, and the implications that this should be understood to have for their autonomy.

Conclusion

The assumption that there is a close relationship between autonomy and rationality in bioethics is well-grounded. Whilst previous theories of rationalist autonomy have made important progress in outlining the kind of role that rationality might play in autonomy, they have been somewhat hampered by certain misunderstandings about the nature of rationality. Furthermore, they have not adequately engaged with the deeper question of why our evaluative judgements should be understood to serve as an appropriate seat of self-government. By drawing on an objectivist account of reasons and the broader literature on philosophy of action in this chapter, I am now in a position to offer the following rationalist minimal conditions of autonomy:

> *Theoretical Rationality*: Decisional autonomy is precluded by theoretically irrational beliefs about information that is material to one's decisions.
> *Practical Rationality*: The autonomous agent's motivating desires must be rational in the following sense:
> They must:
>
> (a) Be endorsed by preferences that are sustained on the basis of the agent's holding (rational) beliefs that, if true, would give the agent reason to pursue the object of the desire.
> And
> (b) These preferences must cohere with other elements of the agent's character system.

In turn, a preference coheres with other elements of an agent's character system if there is a sufficient reason for the agent to maintain that preference in the light of

other competing preferences and theoretically rational acceptances. Coherence is thus incompatible with irrationality, but it is compatible with being less than fully rational.

As I mentioned above, Rebecca Walker claims that a negative rationality condition should replace the standard account's condition concerning the absence of internal control. I agree with this sentiment, but I have made the stronger claim that considerations of practical rationality should feature in a positive condition on autonomy, one that requires that autonomous decisions be grounded by authentic preferences. This account can offer a deeper justification for why practical rationality matters for autonomy.

We may also notice that the positive condition of practical rationality is stronger than the negative criterion of theoretical rationality. One reason for this is that in the case of practical rationality it is possible to draw a meaningful distinction between irrationality and arationality, and both are incompatible with the approach that I am advocating here. The explanation for this is that the positive contribution that practical rationality makes to autonomy is to facilitate our ability to decide in accordance with elements of our character that should be understood to have agential authority. Our decisions can clearly lack that authority if they are irrational, but they also lack it if they are arational. Furthermore, they also lack this authority if they are not endorsed by cohering elements of the agent's character system.

Yet, even if these conditions are necessary they may not be sufficient. It still seems that a suitable theory of autonomy should follow the standard account in maintaining a condition excluding controlling forms of influence such as manipulation, deception, and coercion. It is to these forms of influence that I shall now turn. In particular, in the next chapter, I shall argue that the rationalist conditions that I have set out here can provide a plausible foundation for understanding why manipulation and deception undermine autonomy, and the bearing that this should have on our understanding of authenticity.