Penultimate version. June 2011.
Final version published In L. Radoilska (ed.), *Autonomy and Mental Disorder* (OUP, 2012), pp. ix - xli

Introduction: personal autonomy, decisional capacity, and mental disorder

Lubomira Radoilska

Three premises of the autonomy debate

Autonomy is a fundamental yet contested concept in both philosophy and our broader intellectual culture. To a great extent, this is due to the widely accepted idea that, by giving precedence to reason over tradition, to individuals over communities, autonomy epitomises Enlightenment as an overall project and, more precisely, its core philosophical and political doctrine, liberalism.[1] This genealogy marks the first point of convergence in the current autonomy debate, an illustration of which is the close association between respect for autonomy, on the one hand, and privacy, on the other. For, in both cases, the ambition is to delimit a sphere of individual action which falls beyond the scope of legitimate state authority.[2]

This leads us to a second point of convergence in the current debate, according to which personal autonomy is an individual's right to self-determination, the purpose of which is to protect the exercise

[1] Schneewind (1998) offers a comprehensive analysis of the rise of autonomy as a distinctly modern moral concept. On the close links between autonomy and political liberalism, see Christman and Anderson (2005). A succinct and emblematic outline of the project of Enlightenment is to be found in Kant (1784). See also Horkheimer and Adorno (1973) for a classical critique of this project and its underlying presuppositions.

[2] The relationships between autonomy and authority are helpfully explored in Wolff (1990) and Shapiro (2002). On the significance of a recognised sphere of privacy, where consent between legally competent adults suffices in order to make certain behaviour permissible, see Hart (1963).

of this individual's capacity for self-determination.[3] To simplify, autonomy is an agency concept aiming to define what a person can (legitimately) do. With respect to this second point of convergence, alternative accounts of autonomy could be seen as competing views on the nature of the capacity for self-determination that is worth protecting,[4] whereas critiques of autonomy target the right to self-determination which according to them involves a problematic conceptualisation of freedom in terms of non-interference. In essence, the charge is that, by focusing on *self*-determination, autonomy sets out a misleading ideal of free agency as taking place to the exclusion of others. In doing so, it warrants only a thin 'morality of independence' to the detriment of richer alternatives, such as that of 'mutual responsibility' (Gaylin and Jennings 2003, 4).

A third point of convergence has to do with the assumption that personal autonomy and (severe) mental disorder are mutually exclusive. This is to say that participants in the autonomy debate, who otherwise disagree on both the nature of the capacity for self-determination and the appropriate scope for a right to self-determination, often concur on the idea that mental disorder affects this capacity and, consequently, undercuts the corresponding right. For instance, Marina Oshana articulates this idea as one of the considered intuitions against which rival theories of autonomy are to be assessed. The underlying reasoning is as follows:

[3] The dialectic between a capacity and a right aspect in the idea of autonomy is well articulated in Feinberg (1986, Ch. 18). It also underpins Ronald Dworkin's "integrity view of autonomy", according to which the right to autonomy is meant to protect 'the ability to act out of genuine preference or character or conviction or a sense of self' (1993, 225).

[4] For instance, some of these accounts conceive the capacity at issue in purely psychological terms that, following Christman (2003), can be helpfully divided into two groups, authenticity and competency conditions, the first reflecting the conative and the second the cognitive components of this capacity. In contrast, other accounts point to additional features, which may be psychological or internal as the above, like the capacity to value (Jaworska 1999) but could be also social-relational or external, like being in a position of recognised and secure authority over one's life with respect to powerful others (Oshana 2006).

"It is enough to note that as the possession of these qualities [comprising the capacity for autonomy] is a matter of degree and can be cultivated more or less successfully in persons. A sense of the relevant threshold can be gained by a glance at cases in which *it is clear* that the threshold is not met. The very small child, the individual afflicted with Alzheimer's disease, and the insane person lack the rudimentary ability to self-governing." (Oshana 2006, 7).[5]

To reflect the central role in framing discussions of autonomy, I shall refer to the three points of convergence I identified as 'premises of the autonomy debate'. The first is that autonomy is a fundamentally liberal concept. The second is that autonomy as an agency concept, best understood in terms of a capacity-cum-right to self-determination. The third is that autonomy is incompatible with (severe) mental disorder.

A major ambition of the present collection of essays is to critically examine the third premise and, in so doing, to shed new light onto the preceding two. More precisely, the underlying thought is that, by looking in some detail at cases of mental disorder where autonomy seems to be clearly absent, we are in a better position to articulate the implicit intuitions that underpin our thinking about autonomy and, following this trail, to uncover further conceptual and historical roots that may unsettle some standard assumptions but at the same time offer a clearer perspective onto persistent points of disagreement. A related objective, which this overall strategy aims to achieve, is to help address two kinds of emerging scepticism about autonomy questioning, on the one hand, its theoretical appeal and coherence, and, on the other, its relevance to specific areas of normative thought, including medical and, more specifically, psychiatric ethics.[6]

[5] Both the text inserted in square brackets and the italics are mine.

[6] I consider Nomy Arpaly's suggestion that we should either replace 'autonomy' with narrower, qualified autonomy-concepts, e.g. autonomy as authenticity, autonomy as interpersonal authority etc. etc., or abandon 'autonomy' altogether, for it has become an overworked and misleading term (2003, Ch. 4) as an example of the first kind of scepticism, and Gaylin and Jennings (2003) as an example of the second kind of scepticism.

The notion of decisional capacity is central to the present inquiry, for it is meant to work out the conditions under which the third premise of the autonomy debate obtains, that is, when mental disorder should be considered severe enough as to be incompatible with a capacity for self-determination. Thus, to avoid circularity, it is crucial to get clearer about the relationship between these two capacities. The issue is even more pressing if we consider that 'decisional capacity' is sometimes used as equivalent of 'autonomy' rather than an independent criterion specifying the interactions between autonomy and mental disorder. Moreover, a closer look at the notion of mental disorder itself suggests that this notion builds on the idea of an impediment, if not breakdown of agency, and, in this sense, may depend upon an implicit conception or conceptions of autonomy. To bring into relief the points made so far and set the scene for the following analysis, the next section will look into the right to refuse treatment and its possible limitations. As we shall see, the topic offers a good meeting ground for the concepts at the heart of the present inquiry – personal autonomy, mental disorder, and decisional capacity, – and enables us to explore the close yet challenging links between the three premises of the autonomy debate, with which it began.

Value-neutrality and the capacity threshold

Value-neutrality is a central way of conceptualising autonomy and, possibly, decisional capacity. This is because it offers a plausible interpretation of the conjunction of the first two premises implying that autonomy is an independent but limited in scope source of justification. Drawing on the previous section, autonomy refers to a protected sphere of actions, which we may call purely, or substantively self-regarding and which are permissible solely by virtue of being this person's own, without the need for further justification.[7] In other words, to say that a choice, relative to this sphere of actions, is autonomous is to shield it from legitimate interference, even though this choice may be met with serious objections on moral or prudential grounds. For instance, it could be that this choice harms or disadvantages in some way the person who makes it. Yet, as long as this is a genuine choice as

[7] On the notion of substantively self-regarding choices, see Scoccia (2008).

opposed to being the outcome of coercion, manipulation, or deception and it has no significant negative impact on others than the person who makes this choice, respect for autonomy apparently demands that it remains unopposed. This is the rationale for the so-called Harm principle as a distinctive liberal commitment stating that a person's freedom may be legitimately constrained only to prevent harm to others, but not harm to self, as long as this is willingly and knowingly incurred (Feinberg 1984). As John Stuart Mill, whose work led to formulating the Harm principle observes in his treatise *On Liberty*:

"If a person possesses any tolerable amount of common sense and experience, his own mode of laying out his existence is the best, not because it is the best in itself, but because it is his own mode." (Mill 1859, 114)

In light of these observations, it becomes clear that, by recognising autonomy as an independent source of justification, we are able to efficiently oppose another conception of moral and political obligation, according to which it may sometimes be permissible to interfere with a person's autonomous choice not because it has an identifiable adverse effect on others, but for his or her own good. This conception is known as paternalism. Although it has been recently argued that respect for autonomy and paternalism may not be incompatible after all, it remains the case that, in so far as autonomy is conceived as an independent, though limited in scope source of justification, it does not square with paternalist interventions.[8]

Having said that, there is room for reasonable disagreement about the exact scope of choices where autonomy could provide protection from interference, irrespective of further normative considerations. This point leads to distinguishing, as indicated earlier, substantively self-regarding choices from self-regarding choices which are to a significant degree also other-regarding. As a result,

[8] Examples of this conciliatory approach to autonomy and paternalism include: Scoccia (2008) and Taylor (2004). It builds upon an alternative understanding of autonomy motivated by the rejection of either of the first two premises of the current debate, or an alternative interpretation of their conjunction, which departs from value-neutrality. I shall return to this point in the penultimate section.

some autonomous self-regarding choices may be open to legitimate interference on grounds that they effectively harm unwilling third parties, not only the person who makes the choice at issue (Radoilska 2009). Examples include various health and safety measures, such as the prohibition of smoking in public spaces. This is why instances of treatment refusal are particularly to the point. For it is hard to think of a clearer case where a person's choice should be protected from interference merely by virtue of being his or her own than decisions concerning this person's bodily integrity. A further reason to focus on treatment refusals is that it is often fairly clear that this kind of decision will make the person worse off to the point of causing his or her preventable death. In this respect, they also offer plausible ground for paternalist interventions. To bring out this point, let us consider the following excerpt from a recent court ruling:

"A mentally competent patient has an absolute right to refuse to consent to medical treatment for any reason, rational or irrational, or for no reason at all, even where that decision may lead to his or her own death." (*Re MB* 1997)

This piece of legal reasoning is not only consistent with value-neutrality but effectively parallels the previous quote from *On Liberty* in that it articulates mental competence, also referred to as decisional capacity, as a precondition to this absolute right to treatment refusal. To be more specific, the parallel is with 'any tolerable amount of common sense and experience' conceived as threshold for the application of the 'Harm principle' viz. respect for personal autonomy. This is the third premise of the autonomy debate identified in the previous section.

At first sight, the move from the conjunction of the first two premises understood in terms of value-neutrality to the third premise may seem unproblematic. The reason is as follows. If autonomy is a kind of self-determination, then autonomous choices should be in some sense up to the person who makes them as opposed to merely happening to them (Frankfurt 1971). This idea is crucial to appreciating the underlying argument in *Re MB*. For the conclusion is to confirm that a patient was correctly deemed to lack decisional capacity with respect to giving or refusing consent to a particular treatment – the injection of anaesthetics – because of her extreme phobia of needles. The court decision draws on a distinction between inability to make a specific decision because of a phobia, e.g.

a person cannot make a choice about receiving an injection, on the one hand, and, on the other, an irrational decision to refuse an injection because, for instance, he or she is afraid of needles and yet recognises that needles are not scary. The point of the distinction is to show that only the former, that is, inability to make a decision but not the latter, that is, an irrational decision to the same effect offers a sufficient ground for overturning a patient's explicit treatment refusal.

However, the distinction is not self-evident and, unless we are able to present a further argument to support it, it would seem rather arbitrary. For in both cases, we are faced with apparently similar treatment refusals, that is, made for no good reason and, what is more, in the presence of compelling reasons that speak against it. Following this line of reasoning, if autonomy is an independent source of justification that could shield an irrational treatment refusal, it would seem that it should also be able to shield an incompetent one. Yet, as indicated earlier, if we consider the point of a right to self-determination – to protect a distinctive category of actions merely by virtue of being one's own, irrespective of further considerations, the idea of a threshold satisfied by irrational but not incompetent choices becomes persuasive. This leaves us with a conundrum. The third premise of the autonomy debate appears to be both required by value-neutrality as the conjunction of the first two premises and at the same time at odds with it.

To resolve this conundrum, we should be able to reliably distinguish between irrational treatment refusals that are protected by an absolute right and incompetent treatments refusals that can be overridden on paternalist grounds, i.e. in the patient's best interests. In fact, the court judgment in *Re MB* candidly points to this difficulty in the following observation:

"Although it might be thought that irrationality sits uneasily with competence to decide, panic, indecisiveness and irrationality in themselves do not as such amount to incompetence, but they may be symptoms or evidence of incompetence." (*Re MB* 1997).

A related challenge is to work out a criterion that is not implicitly value-laden, that is, dependent upon further considerations than autonomy itself. Unless this is achieved, some substantively self-regarding choices, such as irrational treatment refusals would turn out to be worthy of respect not merely by virtue of being a person's own, but in so far as they also accord with additional values. Similarly, the

so-called incompetent treatment refusals would be overridden on other grounds than autonomy. But then, a rift between the first two premises or value-neutrality, on the one hand, and the third premise or a threshold condition for autonomy excluding (severe) mental disorder, on the other, will not be avoided. In this case, we may begin to lose sight of what the notion of autonomy, for which we just specified a reliable (value-laden) threshold, even amounts to.

The Mental Capacity Act, which came into power in England and Wales in 2005, as well as related legislation in other countries (Charland 2008), aims to address this challenge by laying down apparently value-neutral criteria for decisional capacity, the threshold condition at issue. Thus, Part 1, section 3 of the Act states that an adult is deemed unable to make a decision for him or herself if he or she cannot at the time of decision-making:

a) understand the information relevant to the decision,

b) retain that information,

c) use or weigh that information as part of the process of making the decision, or

d) communicate his/her decision (whether by talking, using sign language or any other means).

The underlying ambition is to focus on cognitive failures, such as various kinds of misperception of reality and mishandling of evidence, but to exclude any substantive or value-laden criteria. For the sake of clarity I shall refer to the latter kind as 'reasonableness requirements', to distinguish them from the previous minimal or formal criteria consistent with a value-neutral understanding of decisional capacity. How successful has been the Act in keeping out reasonableness requirements?

Looking at conditions **a)** and **c)** above, it is plausible to argue that they both present an implicit reasonableness requirement, especially in light of an explanatory point at the end of section 3. This point reads as follows:

"(4) The information relevant to a decision includes information about the reasonably foreseeable consequences of –

(a) deciding one way or another, or

(b) failing to make the decision." (Mental Capacity Act 2005).

This clarification sits uneasily with one of the principles, set out at the start of the Act, namely: "A person is not to be treated as unable to make a decision merely because he makes an unwise decision." (Mental Capacity Act 2005, 1.1.4). The reason is that if – in order to pass the capacity threshold – a person is expected to 'use or weigh' relevant information, including the 'reasonably foreseeable consequences' specified above in the process of making his or her decision, it becomes difficult to see how this could fail to impose the requirement of making a somewhat wise or at least not particularly unwise decision.

At first sight, this implication may not appear particularly worrying. Yet, accepting it would mean not resolving the difficulty with which we began, that is, how to distinguish between irrational but competent decisions that ought to be protected from interference for the sake of personal autonomy and incompetent ones that ought to be overruled on paternalist grounds. Instead, some if not all irrational decisions would be assimilated to the category of incompetent decisions and become open to paternalist interventions.

To appreciate this point, suffice to look at the way irrationality consistent with decisional capacity is defined by the ruling in *Re MB*:

"Irrationality is here used to connote a decision which is so outrageous in its defiance of logic or of accepted moral standards that no sensible person who had applied his mind to the question to be decided could have arrived at it."

An intuitive reaction is to conclude that this kind of decisions is not worth protecting from interference in the name of autonomy. However, it is precisely in cases like this, where no further moral or prudential considerations speak in favour of an autonomous decision, that the significance of autonomy as an independent source for justification can be assessed. For if we are not prepared to recognise irrational decisions in the sense above as possibly autonomous, we implicitly deny autonomy any greater role than that of a derivative justification. And if so, nothing of consequence would hang on the question whether a choice is autonomous or not, since the right to self-

determination merely stands for a cluster of further substantive rights, like the right to bodily integrity, the right to freedom of religion, etc. etc.[9] By reflecting on the scope of these specific rights, we should be able to determine which instances of treatment refusal are to be honoured and which overridden. From this perspective, the question of whether an irrational treatment refusal could be a person's own in the required sense for autonomy becomes rather tangential. As a result, the point of a distinction between irrational and incompetent decisions is no longer apparent.

It may be tempting to avoid the implication that autonomy is a secondary normative concept by confining scepticism about irrational choices as worth protecting out of respect for autonomy to instances of mental disorder. This solution seems to be in tune with the provisions of the Mental Capacity Act (2005, 1.2.1.), according to which a person lacks capacity with respect to a specific decision, if he or she is unable to make such a decision "because of an impairment of, or a disturbance in the functioning of, the mind or brain". Following this line of reasoning, we could say that, on its own, neither the irrationality of a decision, nor the presence of mental disorder amount to lack of decisional capacity, however, when put together, the two of them add up to it.

Unfortunately, a closer look at this suggestion reveals it as no more than a reformulation of the initial conundrum the conception of decisional capacity was meant to resolve. To recap, the task at hand is to find a criterion that enables us to reliably distinguish, within the category of substantively self-regarding choices of which treatment refusal is a central example, a subcategory of choices that are a person's own in the required sense and are therefore worth protecting merely on grounds of autonomy, irrespective of further normative considerations. Mental disorder comes to attention in this context in so far as it may affect some of a person's motivations to the extent that, with respect to these, he or she is best understood as a 'passive bystander' rather than a self-determining agent (Frankfurt 1971). Clearly, motivations thus affected are not a person's own in the required sense, what is less clear is how to work out a criterion or criteria which could consistently exclude these motivations from the subcategory of substantively self-regarding choices we are interested in. As argued earlier, irrationality cannot play the role of such a criterion on pain of compromising the

[9] See, in particular, Scoccia (2008) advocating such a reductionist account of autonomy.

underlying commitment to autonomy as an independent source of justification. So, if irrational decisions associated with mental disorder are deemed to fall beyond the capacity threshold for autonomy, this cannot be by virtue of their irrationality. The fact that it does make a difference with respect to the issue whether decisions associated with mental disorder are open to paternalist interference or not, should not mislead us. All it shows is that, on this view, no decision associated with mental disorder meets the capacity threshold for autonomy, whether it happens to point toward a rational, or an irrational course of action. This is third premise of the autonomy debate reassessed, but still at odds with the conjunction of the first two by which it is, at the same time, required.

Mental disorder and reasonableness

A possible way out of the persisting conundrum is to reconsider decisional capacity as a threshold for paternalist interventions which only has a partial bearing on the distinction between decisions that should to be protected from interference out of respect for autonomy and decisions that should not. Following this lead, the test for capacity is best understood as involving three separate steps. The object of the first or preliminary step is to answer the question whether a mental disorder that could affect a specific decision is present. To give an example, a diagnosis of depression is prima facie relevant in the context of a life-sustaining treatment refusal, since the wish to end one's life is a core depression symptom. The second or central step is to determine whether the decision at issue is in fact affected by the mental disorder present or not. In terms of the example above, this second step should make it possible to uphold life-sustaining treatment refusals that, although made in the presence of depression, are not influenced by it. By distinguishing between the first two steps of the capacity test, we are able to make sense of the current legislation, according to which the presence of mental disorder is not a bar to the presumption of capacity.

In contrast, the third and final step relates to decisions that in all probability are affected by mental disorder. The point is to make sure that even these decisions do not become open to paternalist interventions by default. Thus, the third step reflects a corollary of the value-neutrality thesis, according to which autonomy is an independent – but not ultimate – source of justification. So, there

might be further moral or prudential reasons that speak in favour of the decisions at issue. In such instances, paternalist interventions still remain out of order although on different grounds than respect for autonomy as conceptualised above.[10]

An immediate advantage of this revision is that it helps make room for some reasonableness requirements which, as observed earlier, seem to be both indispensable for setting out the capacity threshold and difficult to square with the value-neutrality thesis. However, to resolve the conundrum rather than postpone it, the question of reasonableness should only arise at the later stages of the capacity test, ideally at the final and under no circumstances at the preliminary one. To appreciate the significance of the task at hand, let us consider a contrast, which is as intuitive as it is difficult to pin down. This contrast is between life-saving treatment refusals made on religious grounds and refusals with equally fatal consequences, but associated with mental disorder. It may be plausible to want to see only the former but not the latter protected by an absolute right, but does the notion of a capacity threshold support this asymmetry?

A recent *Hastings Center Report* suggests that it may not. By reflecting on the interactions between religious beliefs and decisional capacity in instances of blood transfusion refusals by Jehovah's Witnesses, Adrienne Martin, author of this *Report* concludes:

"Respect for autonomy might require that we respect a treatment decision even when the person 'unreasonably' retains that [religious] belief in spite of compelling counter-evidence – even when the belief renders her incapacited." (Martin 2007, 37).

The idea of unreasonableness here indicates that the treatment refusals at issue are unlikely to pass the capacity threshold if criteria like **a)** and **c)** of the Mental Capacity Act (2005) or equivalents are employed. Such criteria, I called earlier 'reasonableness requirements' aim to ascertain a person's (ability to) use relevant information as part of making the decision under scrutiny. This is not to say that Jehovah's Witnesses are unable to relate relevant treatment refusals back to their core religious

[10] A number of these grounds are covered in section 1.4. 'Best interests' of the Mental Capacity Act (2005). I shall return to this point in the final section of this Introduction.

convictions,[11] but that, unless these convictions are accepted as self-standing premises not to be further probed, the whole reasoning falls apart or, rather, below the capacity threshold. However, isolating religious convictions from assessment in this way seems to defeat the point of a capacity threshold. For if a person's ultimate convictions are to be taken at face value, many decisions based on delusions, a central symptom of severe mental disorder would also pass the test. This is because, in both instances, resistance to counter-evidence and problematic weighing of pros and cons can be explained away assuming certain first premises. No incoherence need be involved.[12] We are now faced with a dilemma: the capacity test is either arbitrary in its application to relevantly similar cases or unfit for purpose, since it fails to isolate even paradigm cases of severe mental disorder.

To avoid this dilemma, it is tempting to dissociate, as Martin (2007) does, the concepts of decisional capacity and personal autonomy. On this ground, it is possible to argue that, whilst life-saving treatment refusals based on religious considerations may fail the capacity test – given that the

[11] Consider, for instance, the following clarification offered by a Jehovah's Witness in the context of a radio debate on the topic of blood transfusion refusals: "Our belief is based firmly on what the Bible has to say, just as the Bible, as everybody knows, forbids things like adultery and stealing and lying, it also tells us to abstain from blood. In a number of places it mentions this command and most especially in Acts chapter 15, in verse 28 and 29 the apostles of Jesus Christ wrote: "For it has seen good to the Holy Spirit and to us to lay upon you no greater burden than these necessary things, that you abstain from what has been sacrificed to idols and from blood and from what is strangled and from unchastity." So on the basis of that we feel the need to obey that command, as well as all the others in the Bible to abstain from things which God says are wrong. The reason that we would refuse blood is because we feel that it would endanger our standing with the Almighty, that it could have an effect on our everlasting future and we're convinced of course that this life is not all that there is, we're convinced that there is a life beyond this and we want to preserve our good relationships with the Almighty so that the judgement he would render would be in our favour." (Parry 2005).

[12] See Jackson (1997) and Jackson and Fulford (1997) exploring the links between religious experience and delusion, and the possibility of non-pathological delusions.

underlying decision-making sometimes clearly frustrates the reasonableness requirements of the test – these refusals should all the same be shielded from interference out of respect for personal autonomy.

However, this move does not resolve the issue of arbitrariness, for nothing tells us why we should not extend the same protection over treatment refusals associated with mental disorder that fail the reasonableness requirements in a relevantly similar way to that of treatment refusals based on religious considerations. And if we pursue this line of thought to its logical conclusion, we are bound to uphold some life-saving treatment refusals that in all probability are affected by mental disorder.

In the abstract, this conclusion could take two forms: either the claim that both categories of treatment refusals pass the capacity test or, alternatively, that neither does, yet both should be respected for the sake of autonomy. However, by accepting the latter form, we would immediately run into the second horn of the dilemma we were trying to avoid, i.e. the capacity test is unfit for purpose. So it is only the former that stands a chance to transcend the dilemma and it requires that, when considering treatment refusals associated with mental disorder, we ignore the kind of unreasonableness they share with treatment refusals made on religious grounds.

A recent decision to uphold a life-saving treatment refusal made by a person with mental health issues, in the aftermath of her tenth suicide attempt (the Wooltorton case) illustrates this logic. Drawing on McLean (2009), the case could be described as follows. Kerrie Wooltorton, a 26-years old woman has ingested antifreeze on nine previous occasions but had accepted life-saving treatment afterwards. She was deemed to have an untreatable emotionally unstable personality disorder and, possibly, to be depressed. In 2007, days before her death, Wooltorton had drafted an advance statement indicating that she did not wish to be treated should the same circumstances arise in the future, even if she called for an ambulance. Rather than being treated, she wanted to die in a situation where she was not alone and where comfort care was provided. A document containing a rejection of treatment was presented by Wooltorton on admission to hospital, after ingesting antifreeze for a tenth time. This document was accepted as valid. In addition, she made a contemporaneous refusal of treatment and was considered to satisfy the criteria for decisional capacity. The medical professionals

involved did not give life-saving treatment. A subsequent coroner's ruling upheld their decision as lawful.[13]

The upshot is that we end up with the same difficulty, with which we began: how to draw a principled and reliable distinction between irrational though competent and incompetent treatment refusals. The several potential resolutions that I considered earlier helped confirm the pivotal role of such a distinction, but at the same time failed to provide the required conceptual means for the task.

To help bring out this point, let us briefly return to the latest suggestion, which was to reinterpret the capacity threshold as related directly to paternalist interventions and loosen its links to personal autonomy. The ambition was to explain the need for some reasonableness requirements when ascertaining decisional capacity without giving up the value-neutrality thesis according to which autonomy is an independent source of justification. Yet, the initially plausible strategy of setting out three separate stages of the capacity test so that the reasonableness requirements are not applied to distinguishing between autonomous and non-autonomous choices but only between competent and incompetent choices was ultimately unsuccessful. More specifically, this strategy led to either ad hoc readjustments where some treatment refusals (e.g. made in the presence of mental disorder) are required to pass a reasonableness test from which others (e.g. made on religious grounds) are exempt, or self-defeating applications of the capacity test, like in the Wooltorton case where treatment refusals that are in all probability due to mental disorder are upheld to avoid ad hocery. In neither case does the permissibility of a treatment refusal rest on respect for personal autonomy as an independent source of justification.

Three promising lines of inquiry

Drawing on the preceding analysis, we are not only able to appreciate the nature and significance of the underlying tension between the three premises of the autonomy debate, and more specifically,

[13] For a further discussion of this case and its implications for clarifying the links between personal autonomy and decisional capacity, see Radoilska (n.d.).

between the value-neutrality thesis jointly supported by the first two premises and the third premise as articulated in the notion of a capacity threshold. In addition, three promising lines of inquiry toward resolving the tension at issue emerge from the discussion.

Firstly, we may abandon the value-neutrality thesis about autonomy. An immediate advantage of this approach is to legitimise the use of reasonableness requirements. In particular, by pointing out specific values and principles, the accord or discord with which makes a choice either autonomous or non-autonomous, it becomes possible to spell out the kinds of unreasonableness that distinguish an incompetent from a (merely) irrational decision. To be successful, accounts following this value-laden line of inquiry should be able to rebut two main charges. The first is of arbitrariness. It builds on the first premise of the autonomy debate, according to which a major task of autonomy is to delimit a recognised sphere of individual action – privacy – outside the scope of legitimate state authority. This task appears to be undermined by reliance on further values or norms, for substantively self-regarding choices which happen not to accord with these values or norms would be exposed to intrusion, irrespective of their private character. The second charge is to miss the point of autonomy as an agency concept (the second premise of the autonomy debate). This charge is particularly pressing for value-laden accounts that either remain silent on the nature of value or take it to be a kind of desire.[14]

[14] More specifically, valuing could be interpreted as desiring to desire (cf. Lewis 1989). There is an intuitive link between this interpretation and the value-neutrality thesis about autonomy. This link is apparent in the so-called hierarchical accounts and in particular Frankfurt (1971) arguing that second-order endorsement of first-order attitudes, e.g. the desire to desire what one actually desires, is the distinctive feature of personal autonomy. See Smith (1992) for a discussion on the connections between understanding valuing as a kind of desire and a Humean psychology according to which all motivation to act is fully accounted for by a belief-desire model where beliefs are taken to be ultimately inert. The distinction between authenticity and competency conditions for autonomy proposed by Christman (2003) reflects this Humean model, see note 4 above. So, by considering values as strong or long-term preferences, value-laden accounts of autonomy and/or decisional

Secondly, we may distinguish, drawing on Williams (1981), between two kinds of reasonableness – internal and external – and argue that the tension between value-neutrality and a capacity threshold only arises when the latter kind is mistakenly brought into the picture along with the former. This is because when assessing whether a decision is internally reasonable, we assume the evaluative perspective of the person who makes the decision. In contrast, when assessing a decision's external reasonableness, we ask ourselves whether there are reasons that speak in favour of it, irrespective of whether the person who makes the decision is able to appreciate these reasons or not.

To get a handle on the proposed distinction, let us briefly return to the issue of life-saving treatment refusals. The difficulty we faced there was how to reliably differentiate refusals that are merely irrational from refusals that are also incompetent. By distinguishing between two separate categories of reasonableness, it becomes possible to isolate irrationality, that is, failures of external reasonableness as irrelevant to establishing the capacity threshold. If only failures of internal reasonableness are inconsistent with decisional capacity, we seem to have found a logical space for the intuitive yet elusive class of irrational but competent life-saving treatment refusals. An example could be a refusal made by a person who considers life to be no longer worth living since changed circumstances made the pursuit of his or her life project impossible. In so far as the latter assessment is correct, the refusal is internally reasonable and therefore passes the capacity threshold. In this context, the question of whether the life project at issue is worthy of such a commitment should not even arise, for this question illegitimately probes the external reasonableness of the treatment refusal.[15]

This approach is clearly attractive, for it promises to satisfy all three premises of the autonomy debate which, although difficult to reconcile, are extremely plausible each in its own right. However, to be successful, accounts pursuing this line of inquiry should be able to maintain a clear distinction between internal and external reasonableness requirements. The underlying difficulty becomes

---

capacity endorse a conception of value and philosophy psychology which arguably undermines their objective (cf. Culver and Gert 2004; Charland 2008).

[15] See Feinberg (1986, 351 – 362) considering a relevantly similar case in some detail.

apparent as soon as we take into consideration the role of interpersonal comparison in (re)constructing the first-person evaluative perspective that frames considerations of internal reasonableness.[16]

Thirdly, we may choose to explore the claims which underpin the premises of the autonomy debate without expecting them to converge into a coherent conceptualisation. More specifically, by assuming a de facto pluralism about the concept(s) of autonomy, we are able to disentangle the variety of intuitions that seem to stand or fall together once we interpret autonomy through the lens of agency (the second premise). Yet, some of these intuitions build on different conceptual grounds. If expressed in terms of agency, their original underpinnings become obfuscated. The upshot is that the credentials of autonomy as an agency concept get undeservedly weakened, since it is evoked as a rationale for a separate group of intuitions. Equally, the intuitions at issue appear as groundless, since they are grafted onto unsuitable roots.

To bring this point to the fore, suffice to look at the three subjects of autonomy attribution as singled out in the *Oxford English Dictionary* (1989) entry on autonomy: institutions (especially states); persons; and organisms. Whilst there is a clear analogy across subjects, the autonomy of each being conceived as independence by virtue of living in accord with one's own laws, the limitations of the analogy are equally clear. This is to say that – by making the kind of autonomy appropriate for one of the subjects above a model for understanding the autonomy of one or both of the remaining two – we would be not merely pushing the analogy too far, but effectively denying the reality of our chosen object of inquiry. For, we would no longer conceive it as governed by its own laws but as following the laws of something or somebody else.

[16] Cf. Davidson (2004, 67): "… in the process of attributing propositional attitudes like beliefs, desires, and preferences to others, interpersonal comparisons are necessarily made. The values that get compared are those of the person who attributes preferences or desires to someone else, and of the person to whom the attributions are made. I do not mean that in attributing a value to another the attributer consciously or unconsciously makes a comparison, but that in the process of attribution the attributer necessarily uses his own values in a way that provides a basis for comparison; a comparison is implied in the attribution."

The salient yet possibly misleading analogy between subjects of autonomy attribution complicates the matter most of all with respect to persons who not only partake in the kind of autonomy specific to them, that is, personal autonomy. As members of states and further self-governing bodies, they play an inherent part in another kind of autonomy, which is political in the broad sense and, as such, attaches primarily to groups and institutions and, only by implication to individuals. As living beings, persons also exhibit a third kind of autonomy, which distinguishes organic entities from artefacts. To have a name for it, let us call it organic autonomy.

In light of these remarks, it becomes apparent that claims which superficially share the following logical form: "We should/ should not treat a person in a particular way out of respect for his/her autonomy", may in fact refer to the political or organic meaning of the term as applied to persons rather than specifically to personal autonomy. This is significant because neither political nor organic autonomy requires a notion of full-blown agency, which however is at the heart of personal autonomy. More precisely, political autonomy builds on an alternative notion, that of full-blown membership within a self-governing community. Seen through this lens, autonomy becomes a status rather than an agency concept.[17] In contrast, organic autonomy draws on a notion of naturalness,

[17] To get clearer on the proposed distinction, we may consider, for instance, Kittay (2005) whose central claim expands on the idea that moral personhood is best defined by the participation in meaningful human relationships instead of the exercise of intricate cognitive capacities allowing one to understand, plan, revise, and, crucially, provide reasons for one's choices and actions. Thus understood, moral personhood is a status, not an agency concept. See also Nussbaum (2009) arguing that no citizen should be excluded from the performance of public functions with deep expressive and symbolic meaning, such as jury service. The ensuing proposal is that people who are unable to do so in person because of, inter alia, severe or profound intellectual disabilities should perform this kind of functions by the proxy of a guardian. Thus understood, citizenship is also a status, not an agency concept. Returning to the autonomy debate, the distinction at issue seems particularly to the point for clarifying the interest and limitations of the so-called relational autonomy, which seems to combine aspects of both a status and an agency concept (esp. Oshana 2006).

according to which living beings are ends in themselves. Unlike the Kantian use of the term, which covers only rational beings whose will is able to give itself its own law, apart from any object willed (*Groundwork*, 4:428), the conceptualisation of living beings as ends in themselves points to a norm of doing well or flourishing that each of them possesses by virtue of substantiating a specific – its own – life form.[18] To treat such a natural end in itself as a mere means to the achievement of an extraneous end is to violate its inherent norm of doing well, that is, its organic autonomy.[19] Following this line of thought, it becomes clear that the impermissibility of harm does not have to be grounded in the infringement of rational agency, e.g. going against someone's self-regarding decisions. This point has direct bearing on both gaining clearer understanding of the Harm principle as key expression of respect for autonomy and locating correctly the difficulty raised by self-harm. For instance, if the impermissibility of harm, as opposed to its (tolerable) wrongness is fully accounted for in terms of infringement of rational agency, self-harm would be obviously permissible since it does not involve such an infringement.[20] Yet, as indicated by the preceding discussion of life-saving treatment refusals, such a view could hardly do justice to the complexities of the issue.

---

[18] The genealogy of the concept of organic autonomy as sketched here can be traced back to Aristotle's definition of nature in *Physics* 2.1. For a recent development, see the notion of natural goodness in Foot (2001, esp. Ch.2).

[19] Some critiques of enhancement vie medical means as opposed to therapy draw on this intuition. See in particular Habermas (2003), whose argument against certain enhancements points to the fact that they obliterate a crucial distinction, that between the 'grown' and the 'made' or, in the terms of the present discussion, a living form and an artefact, brought to existence to serve extraneous ends.

[20] Cf. Dworkin (1994, 156): "…individuals have the right to dispose of their bodily organs and other bodily parts if they so choose. By recognising such a right we respect the bodily autonomy of individuals, that is, their capacity to make choices about how their body is to be treated by others." Dworkin's conception of bodily autonomy not merely differs from that of organic autonomy but is, in a way, a negation of it. More precisely, it translates into the claim that (a certain version of) personal autonomy always trumps the organic autonomy of a person.

So a de facto pluralistic approach to appeals of autonomy could offer the advantage to constructively revisit apparent disagreements on the nature and scope of personal autonomy viz. autonomous agency, as well as explore its possible interactions with the two further notions of autonomy – political and organic – as applicable to persons. With respect to this approach, an obvious drawback to beware of is that, by positing a variety of potentially incompatible autonomy concepts, it may inadvertently slip into a premature scepticism which forecloses rather than advances our systematic understanding of autonomy.

Overview of the chapters

The twelve contributions which compose the present volume explore in detail particular aspects of the three lines of inquiry identified above. More precisely, Chapters 1 – 3, forming the first section "Mapping the conceptual landscape", are focused on key methodological and substantive presuppositions related to personal autonomy, on the one hand, and, mental disorder, on the other. By employing explicit accounts of mental disorder as reference points, Chapters 4 – 6 of the following section "Autonomy in light of mental disorder" critically examine individual premises of the autonomy debate. Chapters 7 – 9 of the penultimate section "Rethinking capacity and respect for autonomy" consider the possible involvement of evaluative commitments both in establishing a capacity threshold and in broadening the scope of respect for autonomy to cover instances where this threshold is unmet. Chapters 10 – 12 of the final section "Emerging alternatives" put forth specific accounts of autonomy with reference to mental disorder. In the following, I shall briefly comment on each contribution in turn.

In "Mental disorder and the value(s) of autonomy" (Chapter 1), Jane Heal identifies and critically examines a form of thought which is implicit in discussions about what we, as a society, owe to people with mental disorder. This form of thought builds upon intuitions which link respect for a person with respect for this person's autonomy. In light of these intuitions, the issue of how to treat a person with mental disorder may seem to revolve around the question whether or not this person has the capacity for autonomy. However, Heal argues, inquiries that share this logical form are

methodologically inappropriate and potentially unhelpful in answering either of the questions they put together: what we owe to people with mental disorder and what is involved in autonomy as a capacity. The reason for this is twofold. Firstly, the apparent consensus about autonomy as a capacity for self-determination that ought to be protected from interference by a corresponding right to self-determination is too shallow to ground a coherent course of action in terms of respect for autonomy. Even if we work with the assumption that autonomy is part of the Enlightenment project, we face an important dilemma since we have to choose between a Kantian or rationality oriented and a Millian or well-being oriented take on the nature and significance of autonomy. Secondly, even if were to reach a substantive consensus on the concept of autonomy, it would arguably require an intricate array of mental capacities, outside the reach of at least some people with mental disorder. Getting clearer on what autonomy is will not help us find out what it means to treat these people respectfully.

In "Autonomy and neuroscience" (Chapter 2), Alfred Mele addresses a sceptical challenge about free will raised by some scientists. This challenge has immediate bearing on the prospects of a coherent notion of personal autonomy. For, if free will is shown to be an illusion as these scientists, following Benjamin Libet, contend, no free actions could ever be performed, yet autonomy as self-determination requires the ability to perform some free actions. How cogent is this kind of scepticism? By reviewing experiments in neuroscience and social psychology cited in its support, Mele concludes that the sceptical thesis about free will rests on a mistake. More precisely, it rests on an unwarranted inference about the alleged insignificance of conscious intentions from the fact that, in certain settings, these intentions seem to be preceded by unconscious preparatory brain activity. Yet, to quote Mele's helpful illustration, "when the lighting of a fuse precedes the burning of the fuse, which in turn precedes a firecracker's exploding, we do not infer that the burning of the fuse plays no causal role in producing the explosion" (p…). Drawing on this analysis, it becomes apparent that the sceptical challenge about free will is in fact motivated by an implicit substance dualism, according to which, in order to be free, our will has to be supernatural. This is because, to be free on this view, a mental action, such as making a decision cannot involve brain events or other physical processes. However, as Mele points out, philosophical arguments in support of free will rarely hang on such an implausible claim. What is

more, there are good reasons to believe that ordinary views about free will are equally independent from substance dualism. Hence, a refutation of the latter does not support scepticism about free will. This conclusion is very much to the point, for some critiques of personal autonomy as an agency concept effectively target substance dualism, which they take to be its metaphysical underpinnings.

In "Three challenges from delusion for theories of autonomy" (Chapter 3), K.W.M. Fulford and I consider a series of issues faced by accounts of autonomy as an agency concept. The central claim is that, to avoid circularity in defining autonomy, these accounts need to explicitly address the putative failures of autonomous agency that are suggested by the logical topography of delusions as paradigm symptoms of mental disorder and, in particular, the possibility of non-pathological or autonomy-preserving delusions. By reflecting on the inescapable yet elusive association between delusions and different kinds of breakdowns of intentional agency that emerges from the variety of case vignettes discussed, we reach two related conclusions. Firstly, there are two separate conceptions of objectivity at work in existing accounts of delusions, only one of which – objectivity as non-arbitrariness – is promising, whereas the other – objectivity as mind-independence – is potentially misleading. Secondly, there is an implicit notion of agential success that underlies our thinking about breakdowns of intentional agency in mental disorder viz. failures of autonomous agency. However, none of the three initially plausible interpretations of agential success that we considered – conventionalist, particularist, and universalist, – could satisfy the legitimate ideal of non-arbitrariness without renouncing the equally intuitive claim that autonomous agency has to do with some credible form of achievement.

As its title suggests, Chapter 4 "Does mental disorder involve loss of personal autonomy?" by Derek Bolton and Natalie Banner is focused on the assumption that mental disorder is incompatible personal autonomy or the third premise of the autonomy debate I identified earlier. More specifically, Bolton and Banner argue that this assumption is best understood if we adopt an experiential account of mental disorder in terms of unmanageable distress. In light of this account, mental disorder leads to different kinds of internal breakdowns of what Bolton and Banner call 'pragmatic autonomy' (p…), i.e. the freedom to do what one is usually able to do. These breakdowns are particularly distressing

not only because of the inability to cope with one's tasks as planned but also because of the resulting disruption of one's self-understanding. This picture may seem familiar from accounts of autonomy which highlight the significance of authenticity and, respectively, consider self-alienation as incompatible with personal autonomy. However, this is a way of thinking about autonomy that Bolton and Banner effectively critique as potentially unhelpful and 'metaphysical', for it seems to stipulate a kind of dualism between a person's real or authentic self and the unauthentic selves, affected by mental disorder. To avoid this unattractive metaphysics of authenticity, Bolton and Banner suggest that we rethink personal autonomy by means of a closer analogy with political autonomy, following recent social-relational accounts.

In "Rationality and self-knowledge in delusion and confabulation: implications for autonomy as self-governance" (Chapter 5), Lisa Bortolotti, Rochelle Cox, Matthew Broome, and Matteo Mameli reflect upon the nature of autonomy as an agency concept or the second premise of the autonomy debate, with which we began. In particular, by exploring the links between irrationality, pathology, and impaired autonomy in delusions and confabulations, Bortolotti and co-authors are able to shed further light on the implicit success criterion identified in Chapter 3. The central question they consider is as follows. Assuming that the failures of epistemic rationality, including self-knowledge involved in the two symptoms of mental disorder above do compromise autonomy, which aspects of it are particularly affected and which – possibly left intact? Drawing on a distinction between autonomy as capacity for self-governance, on the one hand, and autonomy as successful self-governance, on the other, Bortolotti and co-authors argue that delusions and confabulations are not necessarily corruptive of the former kind of autonomy and may even enhance it on occasions; however, they are rarely compatible with the latter kind of autonomy. As Bortolotti et al. put it, "the capacity for self-governance depends on the capacity to develop a self-narrative which encompasses the capacity to endorse attitudes and actions on the basis of reasons. Success in self-governance depends on the coherence of self-narratives and on their correspondence to real life events." (p…). Following this line of thought, it is persuasive to acknowledge that, by allowing a person to strengthen the inner coherence of her self-narrative, some delusions and confabulations, though both irrational and

pathological, could be supportive of her capacity for self-governance. Yet, this support is paradoxical to the extent that more coherence comes at the price of less correspondence to reality and so the seeds of unsuccessful self-governance are already sawn. In other words, delusions and confabulations could help a person make better sense of herself as a planning agent but this would probably affect her getting things right, the other prerequisite of successful planning.

In "Privacy and patient autonomy in mental health care" (Chapter 6), Jennifer Radden argues for a more comprehensive understanding of respect for the autonomy of psychiatric patients, beyond respect for informed consent. More specifically, the discussion is focussed on the 'privacy stakes' for these patients, a term by which Radden refers to the likelihood that their confidentiality will be breached and the degree of harm that they would suffer from it. By reflecting on a series of relevant factors, including: the nature of mental disorder and therapeutic exchange, a normative framework which imposes conflicting professional obligations on mental health care professionals, and the persistent societal stigma associated with mental disorder, Radden concludes that there is a combination of high risk of disclosure and highly negative consequences which place people treated for severe mental disorder "in a situation of extreme and continuing vulnerability" (p…) This upshot has immediate bearing on the issue of personal autonomy. For, by denying a secure right to privacy to psychiatric patients, e.g. on grounds of public safety, the current state of affairs effectively and unfairly compromises their autonomy prospects on recovery.

This line of argument sheds light into a major, yet often neglected side of the dialectic between the capacity for and the right to self-determination, which determines autonomy as a liberal concept at the heart of the Enlightenment project, or the first premise of the autonomy debate outlined at the start of this Introduction. To be more specific, by equating respect for patient autonomy with respect for informed consent, discussions in bioethics frequently focus on one side only of the dialectics at issue, namely, the inference from a capacity to self-determination (in the guise of a capacity to give or refuse consent to medical treatment) to a right to self-determination (in the guise of a right to give or refuse consent to medical treatment). This focus inadvertently overshadows the other side of the autonomy dialectic, according to which a secure right to self-determination viz. a protected sphere of

privacy is a condition of possibility for the capacity for self-determination just as much as it is dependent upon it. The discussion on privacy stakes in mental health care offered by Radden brings to the fore the underlying interdependence between these two sides of personal autonomy.

In "Clarifying Capacity: value and reasons", Jules Holroyd examines the conditions for decisional capacity set out by the Mental Capacity Act (2005) and argues that some of these, namely understanding and weighing relevant information are value-laden. On the one hand, the former condition requires the so-called insight into illness, yet the concepts of health and illness have an implicit evaluative component, such as judging a person to be in a good or a bad state. On the other hand, the latter condition seems to require that a person not only values specific things but also values them to a specific degree relative to others. To illustrate this point, Holroyd looks into capacity assessment in cases of anorexia nervosa, where treatment refusals could be interpreted as undervaluing one's life and well-being relative to thinness because of a distorted or 'pathological' pattern of evaluation. In conclusion, Holroyd points to the need for further discussion in order to resolve some remaining dilemmas about the role of evaluative commitments in current thinking about decisional capacity. As Holroyd observes: "…intuitions seem to pull in different directions: it appears intuitively plausible that over-valuing food avoidance or under-valuing continued existence thwarts the ability to weigh information relevant to treatment decisions. On the other hand it is less intuitively compelling to think that under-valuing the risk of death or disability due to a commitment to religious doctrine undermines decisional capacity" (p…).

The subsequent Chapter 8 "The Mental Capacity Act and conceptions of the good" by Elizabeth Fistein also reflects on the role of evaluative commitments as stipulated by this piece of legislation, however, it focuses on the aspects relevant to decision-making on behalf of people deemed to lack decisional capacity. In particular, Fistein offers a comparative analysis of the ways in which the central notion of best interests is interpreted in theory, law, and clinical practice. According to Fistein, there are three kinds of theories of the good that could underpin this notion: hedonistic, preference satisfaction, and objective list (ideal) theories. Although current legislation, Fistein argues, is moving away from an objective list to a preference satisfaction account of what is in the best interests of a

person lacking capacity, there is evidence to suggest that clinical practice still relies heavily on an objective list theory that prioritises the values of health and safety over a person's known preferences. The case study at the centre of the discussion, which is based on a transcript of clinicians and family members discussing the care of a person, deemed to have lost capacity due to dementia, offers direct insight into the inescapable role played by interpersonal value comparisons in making decisions on behalf of another. Yet, drawing on Fistein's contribution, it is plausible to conclude that this role is not fully recognised by current policy and practice.

In "Autonomy, Value and the First Person" (Chapter 9), Hallvard Lillehammer distinguishes between two separate kinds of autonomy, agent and choice autonomy. The former requires a capacity for substantively self-governing agency as stipulated in the following four necessary conditions: higher order reflection and endorsement of practical options; planning and executing actions that accord with practical options endorsed; responsiveness of the above to minimally intelligible standards of rational argument and; a conception of oneself as a single person living a certain kind of life. On the hand, the latter or choice autonomy is negative freedom with respect to a certain range of options. Unlike agent autonomy, it does not build upon a set of higher-order capacities for rational thought and action. Instead, it is grounded in us having a first-person perspective on life events, a feature at least partly captured by the notion of voluntariness. The distinction between choice and agent autonomy is significant, for it helps identify and forestall a potential confusion about the nature and scope of respect for autonomy, especially with regard to people with severe mental disorders who, ex hypothesi, do not meet all conditions for agent autonomy. The confusion at issue stems from the tempting yet mistaken assumption that choice autonomy is exclusively grounded in agent autonomy. Drawing on this assumption, respect for choice autonomy in the absence of agent autonomy is then either misconceived as respect for a derivative or incomplete form of agent autonomy, or unduly eclipsed by considerations of best interests. The upshot is that human beings are effectively taken to be worthy of respect only to the extent that they are rational agents rather than by virtue of their humanity as expressed in having a first-person perspective on their lives. Yet, as Lillehammer points out: "A human being whose mental capacities does not fully meet all the criteria of genuine self-

governance need not be thought of as a second-rate person any more than a traffic warden need be thought of as a second-rate policeman or an EU citizen claiming residential rights in the UK need be thought of as a second rate Brit." (p…). Along with the line of argument developed in Chapter 1 by Heal, Lillehammer's contribution supports the conclusion that there are good reasons to doubt the centrality of autonomy as an agency concept for specifying what we, as a society, owe to people with severe mental disorders. At the same time, by introducing choice autonomy as a separate category, this contribution points to a plausible alternative for conceptualising respect for autonomy as a value which, though independent of full-blown agency, still functions as a constraint on the promotion of further desirable outcomes.

Drawing on Aristotle's conception of phronesis or practical rationality and, more precisely, its developments in the twentieth-century hermeneutic tradition, "Autonomy, dialogue, and practical rationality" (Chapter 10) by Guy Widdershoven and Tineke Abma offers an account of autonomy, which is centred on moral development as dialogical and practical learning. The central claim is that a focus on freedom from interference is generally unhelpful for conceptualising autonomy and even more so in the context of mental health care. Reflecting on a series of interviews with a person who, having committed a sex offence, was sentenced to treatment in forensic psychiatry, Widdershoven and Abma argue that, by urging patients to reflect on their values through dialogue and joint deliberation, clinicians could help them develop a better practical insight into their situations and, in so doing, promote rather than compromise patient autonomy. Unlike the standard cognitive-oriented approach to autonomy as informed consent, which boils down to providing patients with information whilst leaving their preferences unchallenged, the proposed dialogue-based approach requires that the perspectives of both patients and clinicians are open to challenge and possible transformation as a result of the therapeutic exchange.

In "How do I learn to be me again? Autonomy, life skills, and identity" (Chapter 11), Grant Gillett sets out an account of autonomy as capacity of being-in-the-world-with-others. Drawing on classical works in philosophy and, in particular, Kant's *Critique of Pure Reason* and *Anthropology* which conceptualise insanity as lack of common sense, Gillett argues that mental health recovery hangs on

the ability to think of one's life as meaningful, that is, the ability to share in and have an effect upon the meaning given to the world by others. To develop this ability, a person ought to be actively supported in the exercise of relevant discursive skills so that he or she becomes a moral agent and not merely a 'moral patient' (p…) made to inhabit an unintelligible and frightening intersubjective world. Following this line of reasoning, it becomes apparent that autonomy as ability to exercise control over one's life-world is inseparable from practical rationality as rationality applicable to the ends and not only means of action.

In the concluding Chapter 12 "Autonomy and Ulysses arrangements", I sketch the structure of a general concept of personal autonomy and then reply to possible objections with reference to Ulysses arrangements in psychiatry. The broad lines of this schema are as follows. Unlike the related freedom of action and intentional agency, autonomy is, firstly, incompatible with passive self-determination and, secondly, dependent upon a temporal asymmetry privileging prior over later commitments. More specifically, it takes the form of active self-determination with respect to one's actions, on the one hand, and, on the other, one's motives. There are two ways to exercise active self-determination: trouble-free autonomy and express pre-commitment. The effortlessness that distinguishes the former from the latter makes it difficult to perceive their shared form, which is pre-commitment. In contrast, this comes to light when active self-determination takes place against identifiable threats affecting either a person's authorship (internal obstacles) or ownership (external obstacles) over her actions and motives. The two paradigm kinds of express pre-commitment – trailblazing and character-building – articulate the underlying form, the first with respect to actions, the second with respect to motives.

This analysis points to an implicit hierarchy between three alternative conceptions of autonomy that coexist at present – value-neutral, value-laden, and relational. In particular, a value-neutral approach which conceives autonomy as an independent source of normativity turns out to be central. This is because it covers well the complex relationship of both authorship and ownership over one's actions and motives, at the heart of active self-determination. In contrast, considerations about responsiveness to reasons as opposed to mere incentives, which underpin value-laden conceptions, gain salience only when the presence of significant internal obstacles makes an assumption of non-autonomy plausible.

Similarly, concerns about social-relational status, central to relational conceptions, legitimately come to the fore only when the external obstacles present are so overwhelming as to clearly back an assumption of non-autonomy.

By making explicit the structure of the concept of autonomy, we are in a position to see that the paradox, to which Ulysses arrangements seem to give rise in the context of mental disorder, is in fact due to a flawed conceptualisation that takes effortlessness to be the form of autonomy, not active self-determination. Once this misconception is dispelled, it becomes clear that obstacles to autonomy associated with mental disorder are not different in kind from the obstacles addressed by paradigm instances of express pre-commitment. This is good reason to doubt an assumption of non-autonomy attaching to mental disorder per se.

Acknowledgments

I would like to thank the participants of the following research seminars and conferences at the University of Cambridge: "Autonomy and Mental Health", the Ethics Group, and the Cambridge Forum for Legal and Political Philosophy for many stimulating discussions related to the topics of this volume. I am particularly grateful to: Hallvard Lillehammer, Jane Heal, Ulrich Müller, Matthew Kramer, Jennifer Radden, and Sophia Connell.

I would also like to acknowledge the Wellcome Trust's support for this project (Ref.: 081498/Z/06/Z; 090536MA).

References

Aristotle. (1984). *Physics*, trans. Hardie, R.P. and Gaye, R.K. In Barnes, J. *The Complete Works of Aristotle*. Princeton, N.J.: Princeton University Press.

Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Charland, L. (2008). Decision-Making Capacity. In E. N. Zalta (ed.) The Stanford Encyclopedia of Philosophy (Winter 2008 Edition), URL = < http://plato.stanford.edu/entries/decision-capacity/>

Christman, J. (2003). Autonomy in moral and political philosophy. In E. N. Zalta (ed.) *The Stanford Encyclopaedia of Philosophy* (Winter 2008 Edition), URL = <http://plato.stanford.edu/entries/autonomy-moral/>

Christman, J. and Anderson, J. (eds.) (2005). *Autonomy and the Challenges to Liberalism: New essays*. Cambridge: Cambridge University Press.

Culvert, C.M. and Gert, B. (2004). Competence. In Radden, J. (ed.), *The Philosophy of Psychiatry: A Companion.*Oxford: Oxford University Press; 25 –270.

Davidson, D. (1989). The interpersonal comparison of values. In *Problems of Rationality*. Oxford: Oxford University Press; 59–74.

Dworkin, G. (1994). Markets and morals: the case for organ sales. In Dworkin, G. (ed.), *Morality, Harm, and the Law*. Boulder, Colorado: Westview Press; 155–161.

Dworkin, R. (1993). *Life's Dominion: an Argument about Abortion and Euthanasia*. London: Harper Collins.

Feinberg, J. (1984). *The Moral Limits of the Criminal Law: Vol. 1 Harm to Others*. New York: Oxford University Press.

Feinberg, J. (1986*). The Moral Limits of the Criminal Law: Vol. 3 Harm to Self*. New York: Oxford University Press.

Foot, P. (2001). *Natural Goodness.* Oxford: Clarendon Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Gaylin, W. and Jennings, B. (2003). *The Perversion of Autonomy: Coercion and Constraints in a Liberal Society*. Georgetown University Press.

Habermas, J. (2003). *The Future of Human Nature*. Cambridge: Polity Press

Hart, H.L.A. (1963). *Law, Liberty, and Morality.* Oxford: Oxford University Press.

Horkheimer, M. and Adorno ,T, (1973). *Dialectic of Enlightenment*. London: Allen Lane.

Jackson, M.C. (1997) Benign schizotypy? The case of spiritual experience. In Claridge, G. S.(ed.) *Schizotypy: Relations to Illness and Health*. Oxford: Oxford University Press; 227– 250.

Jackson, M., and Fulford, K.W.M. (1997). Spiritual experience and psychopathology. *Philosophy, Psychiatry, & Psychology* 4(1): 4–66.

Kant, I. (1784). An answer to the question: What is enlightenment? In Kant, I. *Practical Philosophy.* (ed. and trans. Gregor, M.J.) (1996). Cambridge: Cambridge University Press.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals.* In Kant, I. *Practical Philosophy*. (ed. and trans. Gregor, M.J.) (1996). Cambridge: Cambridge University Press.

Kittay, E. (2005). At the margins of moral personhood. *Ethics* 116: 100–131.

Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, suppl. vol. 63: 113–138.

Martin, A. M. (2007). Tales publicly allowed: competence, capacity, and religious belief. *Hastings Center Report* 37 (1): 33–40.

McLean, S. (2009). Live and let die. *British Medical Journal* 339:b4112.

Mental Capacity Act (2005). Office of Public Sector Information, URL = <http://www.opsi.gov.uk/acts/acts2005/ukpga_20050009_en_1>

Mill, J.S. (1859). *On Liberty*. Indianapolis (IN): Bobbs Merrill, 1959.

Nussbaum, M.C. (2009). The capabilities of people with cognitive disabilities. *Metaphilosophy* 40: 331–351.

Parry, V. (2005). Inside the ethics committee: Treating a Jehovah's Witness. BBC Radio 4, 11 May 2005. URL = <http://www.bbc.co.uk/radio4/science/ethicscommittee_20050511.shtml>

Radoilska, L. (2009). Liberalism and public health ethics, *Public Health Ethics* 2(2): 135–145.

Radoilska, L. (n.d.). Autonomy and Depression. In Fulford, K.W.M., Davies, M., Graham, G., Sadler, J., Stanghellini, G. and Thornton, T. (eds.). *Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press. In press.

Re MB [1997] 2 FLR 426, URL = <http://www.bailii.org/ew/cases/EWCA/Civ/1997/3093.html>

Schneewind, J.B. (1998). *The invention of autonomy: a history of modern moral philosophy*. Cambridge: Cambridge University Press.

Scoccia, D. (2008). In defense of hard paternalism. *Law and Philosophy* 27: 351–381.

Shapiro, S. (2002). Authority. In Coleman, J. and Shapiro, S. (eds.). *The Oxford Handbook of Jurisprudence and Philosophy of Law*. Oxford: Oxford University Press; 382–439.

Smith, M. (1992). Valuing: desiring or believing? In Charles, D. and Lennon, K. (eds.), *Reduction, Explanation, and Realism*. Oxford: Clarendon; 323–359.

Taylor, J. S. (2004). Autonomy and informed consent. *Journal of Value Inquiry* 38: 393–391.

Williams, B. (1981). Internal and external reasons. In: *Moral Luck*. Cambridge: Cambridge University Press; 101–113.

Wolff, R.P. (1990). The conflict between authority and autonomy. In Raz. J. (ed.) *Authority*. New York: New York University Press; 20 – 31.