Philosophy & Public Affairs

ORIGINAL ARTICLE OPEN ACCESS

Affect, Autonomy, Authenticity, and the Assessment of Decision-Making Capacity: The Problem of Tyrannical Coherence

Joe Gough[1,2,3]

[1]Philosophy, University of Oxford, Oxford, UK | [2]Merton College, University of Oxford, Oxford, UK | [3]Uehiro Institute, University of Oxford, Oxford, UK

Correspondence: Joe Gough (joe.gough@philosophy.ox.ac.uk; joefgough@gmail.com)

Received: 3 July 2025 | **Revised:** 26 September 2025 | **Accepted:** 31 October 2025

Keywords: authenticity | autonomy | capacity | competence | decisional capacity | decision-making | mental capacity | mental competence

ABSTRACT

There are cases of psychiatric disorder where affective states produce severely self-destructive behavior. Sufferers do not appear to be making autonomous decisions, and appear to be severely impaired in their decision-making capacity. Suffers of these kinds of cases of these kinds of disorders fall into a "gray area" in the law. If this gray area is to be avoided, the law requires clearer criteria for determining how affect can undermine autonomy. Existing "procedural" accounts of autonomy that explicitly set out to deal with how affective states can undermine decision-making are unable to deal with a clinically significant class of such cases. In this class of cases, autonomy is undermined by an affective state that is relevantly coherent with the rest of the person's affective states and attitudes, and has relevantly inoffensive origins. The relevant affective state nevertheless appears to "hijack" the person, and to rule over them "tyrannically." I argue for a necessary condition on autonomy amenable to a procedural account, *non-tyranny*, according to which one is autonomous with respect to a decision only if one has the ability to resist the influence of any given affective state on that decision.

1 | Introduction

A significant area of contemporary philosophical research focuses on understanding how and when people lack decision-making capacity. This research is particularly significant, and of potentially great impact, because the law in many countries, including the United States and the United Kingdom, allows medical practitioners and other state agents to compel treatment and to make decisions on behalf of an individual when they are deemed to lack decision-making capacity (known as "mental capacity" or "mental competence" in law). Bodies of legislation, case law, and professional guidance encode apparently necessary and sufficient conditions for having the capacity to make a decision. Individuals, relative to a particular decision at a particular time, are assessed for whether or not they have decision-making capacity. If they do, their decision must (*ceteris paribus*) be legally respected, no matter how ostensibly unwise it may be. Conversely, if they do not, a decision is taken on their behalf.

The current legal characterization of decision-making capacity is fairly widely considered problematic. The legislation operative in England and Wales, the Mental Capacity Act [1], identifies four necessary and sufficient conditions for being able to make a particular decision at a particular time. One must be able to:

1. Understand the information relevant to the decision;

2. Retain the information for as long as required to make the decision;
3. Use and manipulate that information in the course of making the decision; and
4. Communicate one's decision.[1]

The legislation in many jurisdictions, including Northern Ireland, Scotland, and most of the United States, as well as the practical situation in these jurisdictions, is extremely similar.

One widespread critique of the legislation is that it is too "cognitive" in its focus, giving no role to values and emotions (e.g., [2]). No explicit mention is made of affect, emotion, evaluation, or value. In the courtroom and "in the wild," this produces serious practical issues regarding certain kinds of disorder. There are several kinds of disorder where (a) affective, emotional, motivational, and evaluative states play a significant role, (b) there is no obvious deficit in any of (1–4), (c) decision-making appears to be impaired, and (d) non-intervention, even when it is apparently the expressed wish of the patient, appears to be morally unacceptable because of (c) and the severe self-destructive behavior involved in the disorders. For brevity, from this point on, I shall talk of affective, emotional, motivational, and evaluative states as "affective"—this does not reflect a view of the nature of affectivity; rather, that all four sorts of states give rise to the kind of problem described in (a) to (d), and I suspect that all four sorts of states befit the same kind of solution, described in Section 4.2.

The idea that the kind of decision described in (a) to (d) is not really an autonomous decision made by the sufferer of the disorder has been fleshed out in many ways. Many draw on the notion of "authenticity"—the idea that the decision is somehow not authentic to them, often because it issues from affective states that are not authentic to them. Some further claim that the decision issues instead from the disorder itself, rather than from the person who has the disorder. Many of the key disagreements turn on the nature of authenticity, and what it takes for a state or decision to be authentic to the person.

Others have focused on identifying ways that the affective states involved in these disorders can influence decision-making. For example, they might *circumvent* reflective decisions of the sort adumbrated in criteria (1–4), in the form of strong impulses or irresistible urges. They might *bias* such decisions in ways that might be considered relevant to criterion (3), by making it impossible or harder to consider certain possible outcomes (e.g., [3, 4]; cf. [5, 6]). They might also play motivational roles that circumvent people's reported beliefs (see, e.g., [3, 4]; cf. [5–7]). They might, as such, affect people's interpretations of the world or self-knowledge in ways that are relevant to criteria (1) and (3) [8, 9].

This latter class of account amply demonstrates that the process of decision-making is in fact affected by factors not considered by the legislation. However, it cannot answer the "threshold" question, of when and what kind of influence affects factors on decision-making constitutes incapacity. Some affective states appear to exhibit just the features discussed above, hence *affecting* decision-making, without therefore *undermining* decision-making. Heroic actions stemming from heroic urges in the moment do not appear to reflect undermined decision-making or non-autonomous action, even if they necessarily reflect the circumventing of reflective decision-making. Optimism and pessimism do not generally appear to undermine decision-making, even though they affect probability assignments and consideration of possible outcomes. Self-deception and failures of self-knowledge may often be adaptive and agency enhancing [10], and even where they do not generally appear to undermine the decision-making capacity of the person—if I act badly for reasons I do not acknowledge, this generally appears to be something for which I am rightly held responsible.

We need an account that tells us why some affective factors undermine decision-making by affecting it, but not others. I will aim to produce an account not focused on the *content* of the affective state or the *outcome* of the decisions that it leads to, for several reasons. First, it appears that humans are in fact capable of autonomously deciding to do very stupid and self-destructive things, often for very bad reasons. Second, the legal system has explicitly set itself the goal of having a "procedural" account that identifies a problem in the decision-making process, rather than the outcome or inputs of that process. Finally, most capacity assessments are made primarily by one under-resourced, time-stretched individual: it is clearly undesirable for such an assessor to be deciding what is or is not a good outcome for the person making the decision. The goal of producing a procedural account will be a working assumption of this paper. Many of those who call on content and outcome argue that it is *necessary* to call on such "substantive" considerations because there is no adequate procedural analysis (see endnote 2). I believe that there is a workable procedural account, and hence that this argument is unsound.[2]

Many procedural accounts, which take seriously affective states focus on the notion of authenticity. Accepting that affective states have a heavy impact on decision-making, they aim to say which such states are truly "of" the person, and hence, which decisions issuing from such states are truly "theirs" (see Section 2). There are many such accounts. My contention in this paper is that none of them currently is well-placed to deal with the motivating class of disorders, because of what I call "the problem of tyrannical coherence." The problem of tyrannical coherence is that there are affective states with inoffensive origins and coherent with the rest of one's character that nevertheless appear to undermine autonomy, because they in a sense further specified below hijack and rule tyrannically over the person.

There are nuances here which it is important to get right, regarding my legal intervention. By adding a further criterion for decisional capacity, it may appear that I am narrowing the group of people who would be deemed to be capacitous in service of avoiding "false positives," where people are deemed able to make a decision that they are unable to make. However, things are not so simple. My aim is neither to increase nor to decrease the number of people deemed to lack capacity, but to reduce the amount that clinical and judicial discretion is leaned on to paper over lacunae in the letter of the law.

As Boyle [13] argues, in at least some relevant disorders, it is vanishingly rare that people are deemed to have capacity: he discusses anorexia nervosa, arguing that there are no recorded

legal cases where a person with anorexia nervosa is deemed to have capacity when refusing treatment, and suggesting that people with anorexia nervosa are illicitly assumed to lack decisional capacity—in direct contravention of the aims of capacity legislation to move away from a "status-based" framework and towards a procedural account of decisional capacity. As I argue in Section 5.2, I suspect that this is a direct result of the failure of capacity criteria to unambiguously cover a significant class of cases of anorexia nervosa by the letter of the law. Because the letter of the law does not cover such cases, but in spirit, the law clearly intends to cover such cases, clinical and judicial judgment is used to bridge the gap between the letter and spirit of the law, and paper over a gray area into which many patients fall.

The trouble with having criteria that do not unambiguously cover the relevant class of cases is that it makes clinical and judicial discretion an ineliminable and large part of the implementation of the law. The trouble with introducing such a large element of clinical and judicial discretion is that it provides a mechanism for clinical and judicial biases—in their views of disorders and their sufferers—to affect judgments of capacity even more strongly than they might otherwise. For sufferers of anorexia nervosa, Boyle argues, this results in blanket denials of capacity in treatment refusals. For sufferers of emotionally unstable personality disorder ("EUPD"), patient experience suggests that it results in blanket assertions of capacity often resulting in refusal by doctors to offer treatment [14], and examination of court judgments suggests that the same may be true for sufferers of alcohol use disorder [15]. It plausibly also results in people from non-dominant cultural backgrounds being illicitly denied capacity on the basis of their "alien" values [16]. To reduce the amount of clinical and judicial discretion leaned on in the implementation of the law, and the level of bias-driven false positives and *false negatives* that result from the use of such discretion, clearer criteria for decision-making are required.

2 | Autonomy and Authenticity

Affective states impact decision-making in myriad ways, but only some affective states appear to undermine decision-making capacity and autonomy when they do so. "Authenticity" has come to be used as almost a term of art, such that affective states are "authentic" to a person if and only if they do not undermine that person's decision-making capacity and autonomy when they affect that person's decision-making.

One key distinction among accounts of authenticity, thus understood, is their scope. Some aim only to offer a *sufficient* condition on an affective state's inauthenticity that is relevant to capacity assessments. These more limited accounts are concerned primarily with improving medico-legal assessments of decision-making capacity. Others provide necessary and sufficient conditions on inauthenticity at the most general level. Such accounts often focus on phenomena like manipulation, coercion, and indoctrination.

Another key distinction is between originist and non-originist accounts of authenticity. Originist accounts of authenticity and inauthenticity include criteria regarding the actual historical origin of affective states, while non-originist accounts do not. A particularly well-developed recent non-originist account is Pugh's [17] account. Many non-originist accounts take Frankfurt's [18] account as a starting point, interpreting Frankfurt as claiming that a first-order affective state is "inauthentic" in the relevant sense if it fails to cohere with that person's second-order desires: for example, a desire to do drugs is inauthentic (and undermines decision-making capacity) if one does not also have a desire to desire to do drugs.

Pugh claims that "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" (2020, p. 49). A "preference" for Pugh is "a desire for a certain first-order level desire to be effective in moving the agent to act," and a preference is "personally authorized" when it "coheres" with the agent's other preferences at the time of acting, and with the propositions the agent accepts at the time of acting (pp. 49–50). Authentic affective states, for Pugh, are therefore those that motivate behavior when backed up by a desire for them to motivate behavior that coheres with their other affective states and beliefs.

Among originist accounts, some are "objectivist," bearing directly on the origin of the affective state, while others are "subjectivist," bearing on the attitude that the relevant person would have to the process by which they came to hold the affective state. Some objectivist originist accounts offer a positive criterion—for example, Elster [19] claims that an affective state is authentic if and only if it was reflectively chosen. Others offer a "negative" criterion; for example, Mele [20] claims that an affective state is authentic if and only if the person was *not* compelled to have it.

Christman [21, 22] offers a subjectivist originist account. While Christman's account does make reference to the origin of the relevant affective state, it focuses on the attitude of the relevant person to that origin rather than objective features of that origin. Christman offers three necessary conditions on an affective state's being authentic to the person:

1. Were the person to engage in sustained critical reflection on C over a variety of conditions in the light of the historical processes (adequately described) that gave rise to C;
2. She would not be alienated from C in the sense of feeling and judging that C cannot be sustained as part of an acceptable autobiographical narrative organized by her diachronic practical identity;
3. The reflection being imagined is not constrained by reflection-distorting factors. (2009, p. 154)

That is, an affective state is authentic to a person if they would identify with that affective state given adequate knowledge of and reflection on its origin, unconstrained by reflection-distorting factors.[3]

One key kind of account of authenticity of the narrower sort, focused primarily on medico-legal capacity assessments, is an objectivist originist account. According to this account, an affective state of a person is inauthentic to that person if it is "pathological," in the sense that it is caused by the person's psychiatric disorder [24, 25]. Affective states can be authentic to

the person only if they are non-pathological in this sense. For at least some authors endorsing such a proposal, this was a proposal of last resort, reflecting a lack of faith in the possibility of non-circularly specifying what was wrong with such affective states (Hope, personal communication).

3 | Tyranny

There is a problem—urgently non-hypothetical—for all of these accounts: there are affective states that undermine decision-making capacity and autonomy which are nevertheless coherent with the person's second-order desires, backed up by a personally endorsed preference, reflectively chosen, not compelled, reflectively endorsed and identified with by the person, even in light of knowledge of its origin, and not caused by psychiatric disorder. Indeed, such affective states appear to be relatively widespread at least in cases of eating disorders, major depressive disorder, and addiction.

I will focus on anorexia nervosa in the following, because there are well-developed accounts amenable to my argument. None of what I say in the following is intended to be true of *all* or *most* cases of anorexia nervosa, nor as bearing on the etiology of anorexia nervosa *in general*. Rather, my claims bear on *many* cases, enough to constitute a *significant group* of cases.

Many sufferers of anorexia nervosa are driven by a desire to lose weight. This desire is often one with which they *identify* strongly [7]—a phenomenon called "egosyntonicity." They often have a desire to sustain this desire, and to act on it, both of which cohere with their other affective states and beliefs. Often, before having the disorder, they made a reflective choice to lose weight, one which was not compelled and which they continue to reflectively endorse even in spite of a great deal of forced reflection on that desire and expert testimony and on its history and the fact that they are suffering from a disorder. Because the desire predates the disorder, it cannot plausibly be claimed originally to be caused by the disorder.[4]

This desire to lose weight looks *prima facie* as if it severely undermines the sufferer's autonomy and decision-making. My claim is that previous accounts *miss* a principled way that such a desire to lose weight undermines autonomy and decision-making. The basic idea of my account, which I will flesh out below, is that there are ways that an affective state can become out of control and "tyrannical," without violating any of the requirements on authenticity proposed by the accounts discussed above.

As required by objectivist originist accounts, cases like those described above may start from a desire that is reflectively chosen and not compelled. However, this desire gets completely "out of control" of the sufferer. There are competing models of how this happens, some drawing on habit formation and reward mechanisms, but the general idea is that the desire to lose weight comes to play a disproportionate role in the sufferer's inner life [7]—indeed, this is often the point at which the person is generally viewed as suffering from a disorder. This "disproportionate role" involves several characteristic features, which Charland et al. [26] and Charland [27] use to construe anorexia nervosa as a "passion" in Ribot's sense. It becomes, *inter alia*, highly stable, fixed or obsessional, an organizing scheme for other feelings and emotions, highly motivationally powerful, and highly integrated with cognition and reason.

This out-of-control desire presents a problem for coherentist accounts and subjectivist originist accounts because of the power it exerts over the person's inner life. Coherence of the sort required by Frankfurt and Pugh is achieved, and reflective endorsement of the desire in light of its origin is available, *not* as a reflection of the person's autonomy and the state's authenticity to the person, but *precisely* as a result of how severely the person's autonomy has been undermined—as testament to just how powerful and tyrannical this affective state is within their inner life. When other beliefs or desires, whether first- or second-order, clash with the desire to lose weight, they are revised to fit with the desire to lose weight because of the motivating and organizing power of the desire to lose weight. For example, in extreme cases, when a person suffering from anorexia nervosa is forced to face the incompatibility between their desire to live and their desire to lose weight, they often give up on their desire to live, and say that their desire is to lose weight *even if it kills them*, and that they do not "mind dying" in service of this goal [28, p. 702]. This is the problem of tyrannical coherence: ostensibly, in such cases, an affective state with inoffensive origins and which coheres with the rest of the person's inner life is inauthentic, and severely undermines their autonomy and decision-making.

4 | What is Wrong With Tyranny?

4.1 | Some Previous Accounts

One way to respond to the problem would be to try to claim that this appearance of inoffensive origins, or of coherence, is somehow illusory—and thus to defend existing accounts. I shall assume that it is not acceptable to bite the bullet and claim that patients of the sort described in Section 3 have decision-making capacity, and therefore that their treatment refusals ought to be respected.

One way that the objectivist originist can respond is to claim that the relevant affective state, here, the desire to lose weight, does *not* have inoffensive origins. Young women do not decide to lose weight in a vacuum, but in the face of a society that prizes thinness and weight loss among young women and girls, and additionally, a society that reprimands those who do not live up to societal expectations of this and other kinds. An objectivist originist might therefore be able to claim that this desire is not therefore reflectively chosen [19], or that it is therefore compelled [20].

Either such response casts too many affective states as inauthentic. Religious sentiments, many moral sentiments, many ideas about what constitutes a good life and good life goals, and many banal forms of desire, like one's preferred brand of toothpaste, are socially inculcated in a saliently similar manner. Our views of and attitudes towards the world do not arise ex nihilo. We develop these views and attitudes in a highly social manner [17, 29–33]. The desire to lose weight is especially pernicious because of its *content*, and the *outcome* that successful inculcation

of this desire leads to for many young women and girls, not its *origins*.

Another option for objectivist accounts is to claim that the relevant affective state is *not* the state with inoffensive origins. They might claim that the original desire to lose weight is a distinct affective state from the later desire to lose weight *even if it leads to death*. While the former might have been reflectively chosen, uncompelled, and non-pathological (in the relevant sense; [24]), they might claim that the latter is not. Perhaps the former affective state *causes* the latter affective state, in a way that circumvents reflective choice.

This move appears *ad hoc* unless it is tied to an independently motivated theory of the identity conditions of affective states. I woke up this morning with a desire for toast, but realized that I had run out of bread. I popped out and bought some bread. It feels strange to say that I woke up with a desire for toast, and that on realizing I lacked bread, it caused a distinct desire for toast *even though it meant popping out for bread*. This seems a standard part of inner life, constitutive of having a desire for toast, and not a matter of one affective state causing others. An objectivist originist might claim that the *strength* of a desire is part of its identity conditions. However, in other contexts, it seems more natural to describe affective states—such as my desire to receive fewer emails—as waxing and waning in their strength: the *very same* affective state changes in its strength over time, rather than there being a succession of distinct affective states of different strengths.

Christman's [21, 22] subjectivist originist account might appear to have more promising resources for responding to the problem. His criterion (3) requires that a state can only be authentic when the person's hypothetical reflection on that state is "not constrained by any reflection-distorting factors." The tyrannical desire to lose weight one might claim, is a "reflection-distorting factor." It is far from obvious, however, why this is so on Christman's account. Christman [34] offers several examples of reflection-distorting factors: "the influence of drugs or substances ..., torture or intimidation ..., educational backgrounds ..., and the like." An overwhelming desire to lose weight does not neatly fit with this list of obviously "external" constraints on the person's reflective capacities. If this is the spirit of Christman's proposal, then the problem of tyrannical coherence involves a straightforward counterexample to Christman's account.

However, the idea that the overwhelming desire to lose weight somehow distorts reflection may have an intuitive pull. It is perhaps possible to amend or clarify Christman's account to explain why a tyrannical affective state counts as reflection-distorting. This move makes Christman's account look somewhat redundant in this case. This move simply shifts the problem from explaining why a tyrannical desire is *inauthentic* to explaining why it is *reflection-distorting*, without offering any further resources for answering the problem. Either way, we are faced with the issue of explaining what exactly is going wrong when an affective state comes to rule tyrannically over a person's inner life and motivations.

Frankfurt's account does not have many resources to deal with the problem of tyrannical coherence. Frankfurt [18] has the resources to claim that a person suffering from anorexia nervosa who is recovering, and who no longer wants to lose weight, is not acting autonomously when they are motivated by that desire—but this does not help with earlier stages in the progression of the disorder. Pugh straightforwardly has the resources to claim that in many cases, the person's character system contains incoherence, in that at many stages of the disease people have both a desire to live and a desire to lose weight despite the dangers it poses to them. However, in the highly progressed cases described in Section 3, it is the desire to live that comes out looking inauthentic and incoherent with the majority of the character system on Pugh's account. By the time tyrannical coherence is achieved, no affective states can be ruled inauthentic by Pugh's lights.

Pugh [17] considers a hypothetical case of anorexia nervosa, offering a similar description as I offer in Section 3. He claims that there are two options for his account. In my view, to adopt either strategy is to admit defeat in the attempt to develop a proceduralist account. The first, which he "lean[s] towards," (p. 233) is to bite the bullet, and claim the treatment refusal of such a patient ought to hold legal weight. As stated earlier, a working assumption of this paper is that this is not acceptable, at least if it is possible to give principled reasons why such decisions do not count as autonomous on a proceduralist account.

The second option is to claim that in such cases, the goal of weight loss is assigned "disproportionate strength," [17, p. 232] in a way that reflects a "lack [of] epistemic mechanisms for reliably tracking ... evaluative truths" (p. 225). To assess whether an individual is assigning weight to their goals correctly, in a way that tracks evaluative truths, is to offer a substantive assessment based on the content of their affective state, and hence not a procedural account—despite its appeal to processes for tracking evaluative truths. On this second approach, there is no possibility that someone might autonomously *choose* to prioritize weight loss over all else.

There may be something to be recovered in the *spirit* of the coherence account. It is plausible that Pugh takes coherence within the character system as evidence of occurrent *harmonious* interaction between one's cognitive and affective states—which for him, is reducible to rational interaction. The problem presented by tyrannical coherence is that in such cases, even though there is nothing that can readily be called irrational from a procedural point of view, coherence is evidence *not* of such harmonious interaction, but *rather* of the total hijacking of someone's inner life by a single affective state.

Pugh bites the bullet in claiming that a successfully brainwashed person is autonomous after their brainwashing so long as they have a coherent character system as a result. They simply have a different character system than they had prior to being brainwashed. Pugh [17] considers Mele's [20] example of Beth, a promising philosophy student whose professors enlist neuroscientists to brainwash her to abandon her hobbies and focus all her energy on school. They do so by changing her hierarchies of value, and making her psychologically identical to her fellow philosophy student, Ann. The deep wrong of involuntary brainwashing, in Pugh's view, is that it is deeply wrong for someone's character system to be involuntarily changed in this way.

Pugh may be led to bite the bullet on tyrannical coherence by missing a key difference between it and brainwashing. Beth's values are changed such that she exclusively prioritizes work in philosophy. After brainwashing, she is left psychologically identical to Ann. Assuming that Ann's inner life is harmonious, so too is Beth's newfound inner life. While autonomy-undermining influence was placed on Beth's inner life, no such influence is *occurrently* in place after her brainwashing. Conversely, in cases of tyrannical coherence—if I am correct—there is something *occurrently* wrong with the inner life of the anorexia nervosa sufferer, namely, the inharmonious interaction of an affective state with the rest of her inner life. The difference between the two cases is *not* coherence or the lack thereof, but rather that which coherence is plausibly supposed to evidence: that brainwashed Beth's (newfound) inner states interact appropriately and harmoniously, but the anorexia nervosa sufferer's inner states do not.

4.2 | Tyranny and Control

I have suggested that affective states can be inauthentic in virtue of being out of the person's control, and tyrannical. I have further suggested that this is a matter of internal disharmony and the hijacking of one's inner life. These two characterizations of the problem, one focusing on the relation between the person and the relevant affective state, the other focusing on the relation between the affective state and other inner states, are not offered as identifying different problematic features. Rather, the disharmonious relation between the relevant affective state and the other states of the person is *constitutive* of the deleterious relation between the state and the person as a whole.

Tyrannical affective states exhibit obviously troublesome features that result in a coherent but non-autonomous character system. First, the state is highly *fixed* [26], in that it is not appropriately responsive to other inner states. It sticks around in the face of ostensibly overwhelming considerations, such as its potential lethality and the sufferer's desire to live. It is, in this regard, delusion-like: highly resistant to revision, and not appropriately responsive to the other states of the person.[5] This has led some to propose conceiving of such states as *evaluative delusions* [17, 36].[6]

Second, the state is highly influential. In the cases described, its influence consists of two features. First, *direct motivational force*. People with anorexia nervosa, even once they believe that they should gain weight, and do not want to lose weight, struggle deeply with putting on weight. They struggle to resist the motivational force of the desire to continue to lose weight, a struggle that is exhausting and requires constant strain. Not every desire has this kind of direct motivational force. For example, I genuinely want to be healthy, but when I knowingly do unhealthy things, I do not struggle or feel any significant inner conflict or suffering, nor do I find myself in an exhausting struggle to resist my desire to be healthy.

Another kind of influence that such affective states exert is *inner influence*—the power to change other inner states. In the kinds of cases described in Section 3, when another affective or cognitive state clashes with the desire to lose weight, it is often revised to fit with the desire to lose weight. Not only are such affective states fixed, and not only do they exert direct motivational force, but they also exert a great deal of power over the rest of the person's inner life. It is this feature, combined with their fixity, that results in the coherent character system that poses such a problem for coherentist accounts. Anyone who has quit smoking after being addicted, or to any other addictive substance, is familiar with the phenomenological experience of this inner influence. The most alienating thing about quitting smoking is not that one craves cigarettes despite one's better judgment; instead, it is that these cravings speak to you in a voice that sounds like your own, convincing you that one cigarette cannot hurt, that quitting smoking right now is not that important anyway, or that really you would rather live a shorter life where you smoke. The first-order desire to smoke is accompanied by a significant amount of rationalization, whereby beliefs and desires that clash with this desire, including second-order desires, can be radically changed to fit around the desire to smoke.

At first glance, one might think that fixity is necessary for inauthenticity and the undermining of decision-making and autonomy. The idea might be that if a state is *not* fixed, it can be fairly easily disavowed and eradicated, and that the situation would look radically unlike the non-autonomous cases of anorexia nervosa, major depressive disorder, or addiction in the situation where someone could, with fairly little effort, come not to desire to lose weight, to have low self-worth, to desire death, to desire drugs, and so on. A state with inner influence or direct motivational force, but which could be reflectively rejected and was appropriately sensitive to countervailing considerations, would not obviously undermine autonomy, because it could readily be cut off at the root.

While I do think this helps to show that neither inner influence nor direct motivational force is *sufficient* for inauthenticity, it is fairly easy to generate counterexamples to the idea that fixity is necessary. There are relevant counterexamples where fixity at *no* point seems to play a role. For example, if an intrusive urge with no fixity but with a significant amount of direct motivational force popped into someone's head, and they (for example) kicked a dog impulsively, this does not seem to me to be an autonomous action. The reason for this is fairly obvious: they did not have a chance to reject that urge before it produced action, even though they would have given the chance. Analogous cases are medically relevant. For example, if someone with major depressive disorder declined treatment while under the influence of an intrusive urge to die, one which did not exhibit fixity but did exhibit direct motivational force, this should not be taken as a legally weighty decision for the same reason: the person did not have the chance to reject the urge before it produced action (in this case, saying that they want not to be treated), because it is entirely possible that they would have rejected that urge given the chance.

Fixity also does not appear to be sufficient for inauthenticity. For example, I have, and expect always to have, a fixed desire to smoke. However, it has been a long time since I quit smoking, and it does not exhibit either inner influence or direct motivational force. Were I to act on my desire to smoke, I think it would be an autonomous action because it is currently easy for me not to act on it. It is easy for me not to act on it because it is easy for

me to recognize it as in conflict with my better judgment and broader goals, and because it does not exert direct motivational force on my behavior—it is not an exhausting struggle for me not to smoke; I can simply note that I want to and move on.

Conversely, fixity when combined with *either* direct motivational force, *or* inner influence does seem to be sufficient for inauthenticity—and hence neither direct motivational force nor inner influence seems to be necessary for inauthenticity. For example, if my desire to smoke exhibited direct motivational force that was exhausting to resist, and I smoked when beaten down by the endless battle of resisting, it would not seem to be a fully autonomous action. Likewise, if my desire to smoke exhibited the same kind of inner influence as it did in the early days of quitting, radically reshaping the rest of my inner life to fit around it, it would not seem to be a fully autonomous action were I to take up smoking again.

Pulling these threads together, it seems that a necessary condition on decision-making capacity and autonomy is that the person has the following ability with respect to a given decision

> (*non-tyranny*) to resist the influence of any given affective state on their behavior and decision-making.

Resisting the influence of an affective state can be achieved either by changing it (as can be undermined by fixity or lack of opportunity to change it), or by "isolating" it from one's behavior (as can be undermined by direct motivational force or inner influence). Fixity combined with either inner influence or direct motivational force seems to undermine this ability, and hence to suffice for inauthenticity. Fixity is not necessary, however, because there are situations where an affective state's influence on behavior cannot be resisted even though it does not exhibit fixity—for example, because it appears quickly as an intrusive urge, or because it has already wreaked such havoc on their inner life.[7]

It is not hard to see why non-tyranny might *prima facie* matter to autonomy. It is closely related to the notion of "impaired control" in criminal law [15]. When non-tyranny is not met, and when a tyrannical affective state rules, the person loses *the ability to do otherwise* and therefore their ability to make other choices and their freedom in an at-least-intuitively important sense. Non-tyranny thus accounts for the lack of autonomy exhibited in the kinds of cases discussed in Section 3. Despite the inoffensive origins of the relevant affective state, and despite sufferers' internal coherence, the people are impaired with respect to their decision-making capacity because they are not reasonably able to resist the influence of the relevant affective state on their actions—because that affective state exhibits fixity, inner influence, and direct motivational force. They are firmly "in the grip" of this affective state. It is no longer one factor among many in decision-making, nor one voice among many in one's web of affective states, but rather tyrannically rules over the person in a manner, which undermines their decision-making and autonomy.

Importantly, one could have all of the abilities required by the legislation—to understand, retain, use, and manipulate information, and to communicate one's decision—and yet lack the ability I have called "non-tyranny," and therefore lack decision-making capacity. One could, for example, be capable of understanding both the pathological nature and the consequences of continuing not to eat, retaining that information over the course of making one's decision, using and manipulating that information appropriately as required to inform one's decision by recognizing that it counts against other goals and so on, and yet still be unable to come to any decision other than to continue not to eat, and to be allowed to continue not to eat, because one is *not* able to resist the influence of one's desire on one's decision-making and behavior. This suggests that the legislation fails to offer sufficient conditions for decision-making capacity.

I believe that non-tyranny is a useful criterion for thinking about all four kinds of state that I have grouped together as "affective"—affective, emotional, motivational, and evaluative. One thing that unites these kinds of state is that they are not well covered by existing medico-legal criteria for decision-making, since they are not readily accounted for in terms of information processing, and their effects on decision-making are not readily accounted for in such terms either (e.g., [42–45]). As intimated in the introduction, this does not mean that they never affect information processing (or "intellectual," or "cognitive") capacities, nor that the disorders that involve them do not involve impairments of such capacities (see also Section 5.2). Non-tyranny, however, identifies a way that such states can undermine decision-making *without* any direct effect on information processing capacities. It is an attempt to make precise how non-information-processing states might undermine decisional capacity and autonomy by *circumventing* or *overriding* information processing capacities. A mood such as hopelessness, an emotion such as sadness, an urge or desire, and a value might differ greatly in the precise mechanisms by which they produce behavior. What non-tyranny aims to capture is what it is for such a non-information-processing state to undermine decisional capacity, even understood in the information-processing-focused way that is built into the law: they might produce behavior (by whatever precise mechanism) out of the patient's control, by being such that the patient is *unable to resist* the state's influence on their behavior.

5 | Objections and Clarifications

I have now offered my core proposal. An ability required for one to have the capacity to make a certain decision, and not considered by either the legislation or many previous accounts of autonomy and authenticity, is the ability to resist the influence of any given affective state on that decision. I have called this "non-tyranny." In this final section, I clarify what this amounts to and consider two objections.

5.1 | Clarifications

One important clarification is that my position is not part of a long-since fashionable belief that decision-making is essentially rational, emotions essentially irrational, and hence that emotions and "being emotional" consistently undermine

capacity—that emotions are an irrational embarrassment within the rational mind. It has been suggested to me that perhaps philosophy is currently apologetic about this trend; this is generally good, if it is the case: emotions are essential parts of how we perceive, evaluate, and respond to the world. However, I am convinced that apologetically ignoring the ways that emotions *can* undermine one's decision-making is still failing to take them seriously.

Importantly, the ability to resist the influence of an affective state does not require that one's inner life arises *ex nihilo* in a social vacuum. A person need not be an executive controller or Cartesian ego observing and controlling their own inner life, separate from it. According to several "relational" theories, one develops a self and identity that is authentically one's own through the "conversation" of attitudes, perspectives, and roles originally adopted from the outside world—certain values and views one absorbs can be the basis of criticizing and refining others, and vice versa (see for example, [30–32]). This process is in tension with all aspects of tyranny: a fixed state is immune to such correction and development; a state with inner influence causes the critical conversation to become wholly dominated by a single affective state; and a state with direct motivational force circumvents this conversation.

Another important point is that *being unable to* and *refusing to* resist the influence of an affective state or a value are not the same thing. The former, but not the latter, undermines decision-making capacity. I care strongly about the treatment of animals, and am as such vegan. I do not expect that this will ever change, and I do not expect that I will ever separate this from my behavior. I *refuse* to resist the influence of this affective state on my behavior. Even so, I am *able* to resist its influence—indeed, it would be easy. I do not think I *would* decide to eat meat (e.g., out of politeness, or for health reasons); I do not doubt that I *could* so decide. I also do not doubt that the affective state *could* be changed, although I doubt that it *will* be so changed. Were I presented with a knock-down argument that animals do not suffer, I do not doubt that the affective state *would* change: it is neither fixed, nor does it exhibit inner influence. However, I do not expect that I ever *will be* presented with such an argument. Unlike a tyrannical affective state, it is not merely *coherent* with other parts of my character, but actively reinforced and supported by other parts of my character, including other affective states.

The position is also compatible with the Humean idea that "reason is the slave of the passions"—even if this entails that behavior and decision-making are also subordinated to the passions. Non-tyranny does not require that someone be capable of resisting the influence on their decisions and behavior of the *entire complex* of their affective states. Rather, it requires that no *single* affective state has irresistible influence on their decision-making and behavior. Reason might well be subordinated to the passions, but this does not make it acceptable that reason, behavior, decision-making, and even one's other passions should all be subordinated to a *single* passion. That does not mean that one cannot commit one's life to a single goal or passion—as indicated in the previous paragraph, one can do so autonomously at least when that goal or passion rules not tyrannically, but because it is actively reinforced and supported by one's other attitudes while remaining sensitive to them.

It is also worth stressing that one can gain the ability to resist the influence of an affective state on one's behavior without any intrinsic change to that affective state. Some affective states, "true" compulsions, may be irresistible and tyrannical no matter one's situation and no matter the state of the rest of one's inner ecosystem. However, it is possible for an affective state to be irresistible and tyrannical without being a true compulsion (cf. [38–40]).

This point is worth arguing because Pickard [46–49], in arguing for her "responsibility without blame" framework, appears to presuppose that this is not so. She argues that sufferers of all disorders of the sort I characterized in the introduction, focusing on sufferers of EUPD and addiction, have control of their behavior; hence, they have agency, and hence, they are responsible for their actions (see especially her [48]), even though an affective attitude of blame towards them is inappropriate, by her lights. Such a view is also a significant part of the argument for the abolishment of capacity law in favor of a model of "supported" rather than "substitute" decision-making, as advocated by some—but by no means all—supporters of Article 12 of the UN's Convention of the Rights of Persons with Disabilities ("CRPD"; e.g., [50]; cf. [51]).

In support of this, Pickard highlights several phenomena: that sometimes *believing* that they can change their behavior is sufficient for them to be able to do so; that in a clinical context, practices of "holding to account" such as asking people to explain their motivations, encouraging them to act otherwise, and imposing negative consequences are commonplace; that clinical communities encourage members to see themselves as responsible [46–48]; that the sufferers of these disorders are sometimes capable of motivational trade-offs (e.g., choosing money over drugs; [48, 49]); and that many addicts "mature out" of their addictions as they acquire more goals, responsibilities, and opportunities for enjoyment [49]. She claims that the "natural explanation" of such phenomena is that sufferers *can* "exercise the willpower necessary," but only do so when they are sufficiently motivated [48, p. 145]. She further claims that many forms of talk therapy aim to help patients to "engage" their capacity to control their behavior, hence presupposing that they are able to do so [47, p. 1137]. She encourages skepticism, on this basis, of the testimony of sufferers of these disorders who deny that they have, and had, control over their pathological behaviors [48, p. 147].

However, it follows from none of these considerations that sufferers of such disorders are able to resist the influence of the relevant affective states on their decision-making behavior. It might instead be the case that one can *gain* the ability to resist the influence of an affective state through the kinds of change and intervention that Pickard points to. It seems to me far more "natural" to claim that sufferers of these disorders *gain* the ability to exercise the willpower necessary to change their behavior through such changes, than to claim that they had it all along but chose not to exercise it because they were not motivated to do so. That it is possible for patients to engage their capacity for choice and control *with training, support, and encouragement* does not mean that they presuppose that patients are able to do so *without such training, support, or encouragement*. One might be unable to resist the influence of a fixed affective state with

direct motivational force *precisely because* one, for example, believes that one cannot, has no countervailing affective states, or lacks the skills required to engage one's capacities for behavioral control. My father, like many men of his generation, smoked from the age of 12 and quit when he had children. He wanted to quit for many years before that, but was unable to do so. It seems wrong to claim that the countervailing force of the sense that he might be endangering his children gave him the motivation to choose to exercise willpower that he had all along. Instead, this new affective state gave him *more* willpower, enough to resist the force of the urge to smoke, which he was previously unable to resist.

5.2 | Reform Unneeded

Sections 4.21 and 4.22 of the Mental Capacity Act Code of Practice [52] elaborate criterion (3) to suggest that certain kinds of affective state can undermine decision-making because they constitute considerations "too strong to ignore." This proposal is close in spirit to my own, plausibly intended to deal with the same issues as I am raising here. One might therefore object to my proposal that it is unneeded, and that the legislation already has the tools required to deal with tyrannical coherence.

This objection fails because while perhaps *intended* to cover the same issues, this elaboration of criterion 3 does not offer an equivalent condition to non-tyranny, and in my view, does not succeed in dealing with the motivating issues. Being unable to *ignore* a certain affective state is by no means sufficient for tyranny, and for being unable to *resist its influence* on one's decision-making and behavior. Ignoring an affective state is by no means necessary as part of resisting its influence. One might be unable to ignore an affective state, and yet still be *able* to consider countervailing considerations, still be *able* to change that state should strong countervailing considerations come to light, and still be *able* to isolate that state from one's behavior—even in cases where one does not in fact do so.

The code of practice was written alongside the Mental Capacity Act, to guide its use especially in situations where it is hard to apply. Sections 4.21 and 4.22 are written in part to deal with the application of the Act to the class of disorders under consideration in this paper, specifically mentioning the "compulsion not to eat" in anorexia nervosa. These sections attempt to gerrymander considerations of affect into criterion (3). The key philosophical problem is pressed by Pugh [17]: many considerations are impossible to ignore, and yet do not undermine autonomy or decision-making. He considers the example of Sue, a woman with extraordinarily mild recurring headaches, who is told about a surgery that might cure her headaches, but would have a 90% chance of killing her. He claims that the mortality rate, presumably combined with her wish not to die, is a consideration too strong to ignore. Even so, it patently does not thereby undermine her autonomy or decision-making.[8]

This flawed attempt has produced other problems. These sections encourage the conflation of one *not* using certain information (because of overwhelming considerations) with one *not being able to* use that information (as encoded in the original legislation). This conflation represents an attempt at a quick fix, in my view, attempting to account for tyranny in information-processing terms, by claiming that what is wrong in cases of tyranny is that one does not use all the information available in making one's decision. The conflation has to be made because in many such cases, the sufferer's information processing capacities are intact, and hence they appear to be *able* to use that information.

Even if in theory, non-tyranny *could* somehow be subsumed under criterion (3), there is a clear need for reform of the legislation and code of practice in England and Wales. The current situation of capacity assessments "in the wild" where affective states are highly relevant is, frankly, nightmarish. The legislation, read narrowly, would lead to many false positives, where patients were ruled to have the capacity to make a decision despite lacking it. Read broadly, and taking the code of practice seriously, it would result in many false negatives, where incapacity was found simply on the basis that there were *any* overwhelming considerations involved in the decision.

Discretion is the key tool psychiatrists use to thread the needle between the false positives that would result from focusing on the legislation alone, and the false negatives that would result from taking the code of practice seriously and allowing that the presence of any overwhelming considerations can undermine decision-making capacity. Psychiatrists are forced by the inadequate legislation to read the legislation as broadly or narrowly as required to support their own decisions about treatment, based on their own views of the right outcome.

This results in pernicious kinds of false positives and false negatives that reflect the biases of the people who actually make capacity assessments. As Boyle [13] argues, patients with anorexia nervosa are consistently denied capacity in a way that does not plausibly reflect a procedural approach to capacity. Craigie and Davies [15] argue that the MCA, read together with the code of practice, is sometimes read broadly enough by clinicians to qualify as an "impaired control" test of the sort used in criminal law and closely related to non-tyranny. However, they also show that clinicians and judges do so for some disorders but not others. The test is read narrowly in relation to forms of addiction, such as alcohol dependence, which are therefore generally not taken to undermine capacity, and broadly in relation to anorexia. Craigie and Davies argue that this reflects insidious value judgments about the worthiness and virtuousness of the sufferers of these disorders:

> What strongly distinguishes substance dependency from anorexia is that anorexia develops in the pursuit of things that our society values. Thinness and exercise are considered virtuous, along with traits associated with anorexia such as perfectionism and resistance against bodily desires. (p. 238).

Some pernicious false positives involve so-called "difficult patients" and disorders such as EUPD, which bestow a patient with "difficult patient" status [46]. People with EUPD perform well on general cognitive tests, and appear not to lack any of the criteria in the legislation. A widely attested phenomenon among people with EUPD in the United Kingdom is that

they turn up to a hospital asking for help with overwhelming suicidal urges. They are given the standard battery of tests, and deemed to have capacity. They are, as such, denied admission—they are told that they have the capacity to decide whether or not to kill themselves, and hence are in no urgent need of medical assistance [14]. In cases such as these, the discretion afforded to clinicians allows them to justify not helping people with urgent, warranted worries that they will kill themselves incapacitously.

Some pernicious false negatives involve people who do not hold or weight values in the culturally dominant manner. A particularly pernicious kind of case involves people with non-dominant cultural backgrounds. Although the issue remains unstudied, and hence it is not possible to make empirical claims, many suspect that people with non-dominant cultural backgrounds are deemed to lack capacity at higher rates than people whose upbringing reflects the dominant culture—just as they are deemed to be of greater risk to themselves and others, and detained, at higher rates. This issue, and the dearth of research on it, has been highlighted by Dr. Shubulade Smith CBE, the first black president of the Royal College of Psychiatrists, in an interview with a leading mental capacity lawyer, Alex Ruck Keene KC (Hon) ([16]). The issue stems from the fact that the values, goals, and desires of these people are more often "alien" to the person assessing their decision-making capacity. The worry is that they are deemed to assign excessive weight to certain values, and hence to lack the capacity to make the relevant treatment decision—exactly the same rationale as is often used in the case of patients with anorexia nervosa.

Non-tyranny helps to identify what is going wrong in these kinds of cases. In the case of people with EUPD, while their information-processing capacities are intact, their suicidal urges come on suddenly and exhibit a great deal of direct motivational force (as well, often, as inner influence and fixity). It is for this reason that they are not, in fact, capacitious in deciding whether or not to take their own lives. In the case of people from non-dominant cultural backgrounds, the fact that they hold different and differently weighted values than the assessor, in a way the assessor views as incorrect, is simply not relevant to whether or not they are capacitious. Nor is it relevant that their values lead them not to use information that the assessor views as relevant in making their decision about treatment. The relevant questions are whether they are *able* to resist the influence of those values on their behavior, and whether or not they are *able* to consider the other relevant information.[9]

There is another form of argument available that reform is unneeded. According to this argument, the law is already adequate because there are cognitive deficits associated with the relevant disorders, and hence there is no need to add explicit consideration of non-cognitive factors into the law. Relevant cognitive deficits have been identified in many relevant disorders, including anorexia nervosa [13], depression ([53]; Owen et al. 2015 [54]), psychosis, and mania [55]. This is also tied to a relevant worry about my central example, anorexia nervosa: one might worry that it is not an analytically clean example, since anorexia nervosa also has significant physiological effects as the body starves, which in turn may result in cognitive impairments.

This is valuable work, and while I do not think that it renders my proposal unmotivated, I am keen for my proposal to be seen as complementary to rather than in competition with such work. I do not, and need not, deny that such cognitive deficits exist, are significant to questions of capacity, and ought to be considered more in actual assessments of capacity. My proposal is complementary to such proposals in that (a) the relevant cognitive deficits are not—in any of the above-cited papers—argued to be universal in the relevant disorders, and (b) it is a very real possibility that there are cases of the relevant disorders, which undermine capacity without the presence of such deficits. My proposal offers a principled procedural criterion for thinking about how a case of one of the relevant disorders might undermine capacity in the absence of such cognitive deficits—whether these are directly associated with the disorder, or caused by means of physiological effects of the disorder.

For example, Owen et al. [54] argue that cognitive temporal inabilities are associated with many cases of depression, in particular, a class of "severe" cases. I can attest to this based on my own experience of depression. However, even when my depression manifests in a milder form and is not so severe as to result in such temporal inabilities, it can result in strong suicidal impulses. If I were called on to make an urgent treatment decision while experiencing such an urge, I think it likely that I would decline treatment, and that I would not be capacitous when I did so. The reason that I think that I would not be capacitous in declining treatment is that (i) my suicidal urges have direct motivational force of a sort that is extremely hard to resist, especially in situations of stress, and (ii) in the context of an urgent treatment decision, I would not have the opportunity to resist, and it would not be practicable for the assessor to wait for the urge to pass. That is, proposals that identify cognitive deficits identify one way (already covered by the legislation) that the relevant disorders can undermine capacity; my proposal is complementary to such proposals in identifying a *distinct* way that they can undermine capacity.

A final argument against reform is that it is in tension with the movement towards greater recognition of the legal capacity of persons with disabilities embodied in the UN's CRPD, especially its Article 12. Addressing this in detail would take me far beyond the intended scope of this paper. Some, but by no means all or even most, of the proponents of the CRPD advocate for the abolition of capacity law, and indeed, for the abolition of all forms of non-consensual medical treatment (e.g., [50]). Many of the proponents of the CRPD, however, offer a less extreme conclusion, in no small part because an extreme version of this position would make it illegal for paramedics to treat unconscious patients—such treatment would, after all, be non-consensual [56]. The European Court of Human Rights, for example, has concluded that the CRPD is consistent with the retention not only of capacity law, but also of guardianship regimes [57].

A uniting ethos of most proponents of the CRPD is to empower people to make decisions for themselves whenever possible, by the provision of support, perhaps alongside a limited form of substitute decision-making [58]. The reform I am advocating is still relevant, even in the move towards more supported, and less substitute, decision-making. It is relevant if, as some argue, it is likely that the same "threshold questions" will be

reproduced under the new regime (e.g., the central question may become whose decisions *made without support* should be respected; [51]). It is also relevant because supported decision-making is often taken largely to be a matter of eliciting and enacting people's "will and preferences," in light of their rights. As such, many authors in this literature have devoted significant attention to distinguishing "pathological" from "non-pathological," or autonomy-facilitating from autonomy-hindering, affective states, since there are frequently clashes among and between one's will, preferences, and rights (e.g., [59–61]).

It is worth conceding that non-tyranny does not make the assessment of capacity *easy*. It is hard to assess whether someone is unable to resist a given affective state, especially in light of the incentives that they may feel to be dishonest about whether they are able to resist it—to claim that they can resist it, when they in fact cannot. It is worth noting, on this point, that none of the criteria already included as part of the construct of decisional capacity are easy to assess in every case. I have been told of patients reporting abuse from their relatives which, if delusional, would be relevant to criterion 1; in the absence of resources for information-gathering, this leaves assessors guessing as to the truth. Criterion 3 is famously difficult to assess, in the absence of widely-agreed standards for using and manipulating information.

One might think the situation is still different from the case of non-tyranny. In particular, one might worry that the legislation does not include a criterion akin to non-tyranny because there are no tools for measuring it, unlike the information-processing abilities included in the legislation (cf, e.g., [62]). This is not so. There exist psychometric tools for measuring relevant constructs, *emotional competence* and *emotional regulation*, already shown to be valid and predictive in the context of the relevant class of disorders [63, 64]. Indeed, all of the relevant disorders are associated with deficits in emotional regulation. Such tools cannot directly be used to measure whether a particular affective state is tyrannical—but nor can the kinds of general cognitive tests currently used in capacity assessments establish whether a person is capable of understanding, retaining, using or manipulating particular pieces of information. Rather, they have an evidential relationship to the ultimate determination.

Even though it may not be easy to answer the question of non-tyranny, it is nevertheless necessary. Its necessity is indicated by the "considerations too strong to ignore" criterion built into criterion 3 by the Code of Practice. It is no easier to assess whether there are considerations the patient is unable to ignore than to assess non-tyranny. It is, more importantly, the wrong question, as argued above. If it were the case that a criterion like non-tyranny had been left out of the legislation because it would be hard to assess, then the reality of the law as implemented suggests that this was short-sighted: asking the wrong question, even if it is easier, is not a good idea, and has resulted in the kind of bias-heavy discretion in judgments of capacity discussed above. It may be hard to establish whether patients are in fact able to resist a given affective state or not, but this should be the goal of the clinician's questioning, and the goal of judicial reasoning. We are living in the alternative.

A particular version of this worry is that clinicians will indulge in circular reasoning: that they will conclude that a patient lacks capacity because their decision is characteristic of the disorder, and hence evidence that they are unable to resist the influence of the affective states associated with the disorder—for example, that a patient with anorexia nervosa lacks capacity because they are choosing not to eat, and choosing not to eat is characteristic of anorexia nervosa, hence evidence that they are unable to resist the desire not to eat, as associated with anorexia nervosa. If this worry is offered as an objection to my proposal, the trouble is that we are *already* in a situation where this kind of circular reasoning is indulged in [13]. The "considerations too strong to ignore" do not guard against this kind of circular reasoning, and offers more tools to justify it, since a huge number of decisions involve a "consideration too strong to ignore." Experiencing a "consideration too strong to ignore" also does not suffice for tyranny. Compared with the existing legislation (as implemented), the non-tyranny requirement provides fewer tools for justifying this kind of circular reasoning, since it requires assessors to present evidence that the patient is unable to *resist the influence of* affective states on their behavior, rather than merely being unable to *ignore* those affective states.

5.3 | Other Tyrannical States: The Will to Live

Another way to object to my proposal is to point towards counterexamples—affective states that exhibit tyranny yet do not appear to be inauthentic and undermine decision-making or autonomy. I can think of at least one plausible example of a state that exhibits tyranny and yet is widely accepted and might perhaps not be seen as undermining autonomy.

The will to live is extremely hard to resist, and extremely hard to lose. Even people who are suicidal are claimed to exhibit a great deal of ambivalence, and to need to "train" themselves to resist the force of their will to live in order to kill themselves [65–68]. It therefore appears to exhibit at least fixity and direct motivational force. It is plausible that, for many people, it also exerts inner influence—that they work to make their other beliefs and affective states coherent with their will to live, for example, by believing that life is worth living. I will accept that, when exhibiting these features, the will to live counts as tyrannical.

What I deny is that the will to live is *not*, therefore, inauthentic and autonomy-undermining when it exhibits these features. This might sound initially strange, but consider cases of coercion. Coercion in the form of threats to one's life can undermine one's decision-making and autonomy, and result in decisions that are not fully autonomous. Why? One straightforward account is that the will to live is, in such situations, tyrannical, and largely responsible for the decisions one makes in such situations of coercion. If one, for example, betrays a friend because one is coerced to do so by a threat to one's life, a simple account of why this does not reflect an autonomous decision is that one was not able to resist the influence of one's will to live in producing that betrayal.

Importantly, this does not mean that *no* autonomous decisions issue from one's will to live—even if it is fixed. There are situations where people choose to sacrifice themselves and their

lives for a cause—showing that it is, at least in some contexts, for some people, possible to resist the influence of one's will to live. If such a person, in such a context, chose *not* to sacrifice their lives because of their will to live, I think that this would count as an autonomous decision—they would be choosing not to sacrifice their lives because of their will to live, even though they *could* have resisted the influence of their will to live.

The situation can perhaps get murky in ways that are important for medico-legal purposes. Consider, for example, the person who has for some time had a long-standing, reflective preference for cancer treatment to be withdrawn, and who thinks that it is time to give up on acting on their will to live. They might, in the moment, be driven by their will to live to balk and *not* to refuse treatment. This decision would, by my lights, not be autonomous: despite their reflective preferences and overall views, they were unable to resist the impact of their will to live.

I have spoken to some who think it would nevertheless be wrong *not* to treat such a patient. I admit that I do not feel the pull of this view. Even so, granting for the sake of argument that it would be wrong not to treat them, this only constitutes a counterexample to my account if the *reason* that it is wrong not to treat them is that their decision is in fact *autonomous*, as my account denies. However, this is not obviously the case. Even assuming that it is wrong not to treat them, this may be because it is in their *best interest* to treat them. Legally, if someone is ruled incapacitous to make a decision, the relevant state actor is empowered to judge their best interest. Any intuition that it is wrong not to treat them, in such a case, may stem from the fact that even though their apparent decision to accept treatment is not a capacitous one, it is nevertheless *prima facie* in their best interest to treat them.

Several people have also mentioned to me that one's love for one's children may likewise count as undermining autonomy on my account. I believe that this is correct—plausibly, if one sees one's child in danger, one *cannot help but* move to help them, because one is unable to resist the influence of one's love on one's behavior. I do not, however, think that this is a counterintuitive result. First, it does not mean that others should be empowered to intervene: one's "best interest," values, and life goals may well align with the (by my lights) non-autonomous actions that result. Moreover, it seems to be quite widely acknowledged, and often seen as admirable, that we may be driven by our love for others to "good" actions, which are *outside our control*: that we, as I put it above, *cannot help but* take certain actions in the interests of our loved ones; indeed, many seem to believe that we *should not be able to* resist taking some actions out of love. My proposal may widen the net on which affective states undermine our decisional capacity and autonomy compared with some previous proposals. However, not all the actions that result from affective states that undermine our autonomy seem to be undesirable, and many seem to be part of a good life. I do not mind biting this bullet: autonomy is not the sole ingredient for a good life. It seems entirely plausible to me that many of us give up some amount of autonomy in service of our love for others, and that for many of us, this is part of the best life possible for us.

6 | Conclusion

Motivated by a clinically significant class of examples, I have offered a procedural criterion that aims to account for the problem of tyrannical coherence—*non-tyranny*, the ability to resist the influence of any given affective state on one's behavior and decision-making. This proposal, if it is tenable, appears to indicate that it is not necessary to abandon the attempt to offer a procedural account of what is going wrong in such cases in favor of an account based on substantive considerations of the content of affective states and the outcomes of decisions. However, it also appears to indicate that it is necessary to explicitly consider how affective states interact with other states in decision-making in offering an adequate procedural account of decision-making capacity, authenticity, and autonomy, and to consider the ways that affective states ought to be open to correction and isolable from behavior.

Acknowledgments

The author's thanks to Sarah Wieten, Emily Webster, Helen Gough, David Faraci, Fintan Mallory, Nick Upton, Jonathan Pugh, Anna Golova, Katie Zhou, Stephen Gadsby, the attendees of Matt Parrott's *Dynamics of Delusions* seminar at the University of Oxford, and the attendees of a presentation at the 2024 meeting of the European Society of Philosophy and Psychology for feedback on earlier versions of this piece. Special thanks to two anonymous reviewers of this journal. With thanks to the British Academy for funding the project of which this is a part.

Funding

This work was supported by the British Academy.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Endnotes

[1] The last of these is not generally considered an internal part of the decision-making process; rather, it has played an important practical purpose, namely, bringing people who are *not* able to communicate their decision (specifically, people with locked-in syndrome) under the legal protection of the best-interest substitute decision-making framework.

[2] Several accounts of autonomy and decision-making, including in the context of mental capacity law and psychiatric illness, argue that "substantive" considerations of content and outcome are necessary, because "formal" or "procedural" accounts are insufficient. Many such accounts describe themselves as "relational" (e.g., [11, 12]), and are explicitly feminist in their outlook and the body of theory they draw on. My aim in the following is by no means to directly argue against such accounts. My aim is to generate a better "procedural" account, and hence they are not directly relevant to my argument. In a sense, therefore, my argument is conditional on the correct account of autonomy for medico-legal purposes being procedural. Nevertheless, in as much as the kinds of case I consider are motivation for such accounts, and in as much as I can show that such cases can be satisfactorily dealt with by a procedural account, my argument undercuts the motivation for these accounts.

[3] Dworkin [23] offers an originist account with both subjectivist and objectivist elements.

[4] One could amend the pathology account not to be originist, and claim that the desire is inauthentic because although not *originally* caused by the disorder, it is *causally maintained* by it. This is not against the spirit of my own account. However, given the controversies surrounding the nature of psychiatric pathology and its relation to symptoms, it is not a very informative claim. In my view, the best way to flesh out this claim is in line with my own account—that is, to claim that meeting the conditions I lay out in Section 4.2 is part of what it is for a state to be pathologically maintained.

[5] A caveat is in order here regarding the notion of fixity: fixity does not consist in absolute unchangeability, but rather in the extremely high resistance of a state to responding appropriately to countervailing considerations. First, some delusions—and some affective states of the sort I am concerned with, such as the wish to die in major depressive disorder—ebb and flow [5, 6]. Even so, the point remains that they are not appropriately responsive to the person's other states, rather, they do so seemingly of their own accord. Second, some delusions—and some relevant affective states—are not *totally* unrevisable or unresponsive to countervailing considerations and evidence [35]. Even so, they are highly *resistant* to revision and to responding appropriately to countervailing considerations and evidence [35].

[6] I will not use the term "evaluative delusion" in the following, because I do not believe it clarifies matters. While I believe that the analogy between such affective states and delusions is illuminating, I believe that the notion of evaluative delusions raises as many questions as it answers. Assuming that delusions are beliefs [37], there appear to be at least four ways of construing the relationship between the relevant affective state and the relevant evaluative belief: the belief might cause the affective state, or it might manifest as the affective state; the affective state might cause the evaluative belief, or it might somehow instantiate or "count as" that evaluative belief. The former two, "belief-first" construals of the relation do not fit anorexia, where the desire to lose weight appears to predate later evaluative beliefs about the importance of doing so [26]. The latter two "affect-first" construals of the relation fit better with the discussion of anorexia, but such "affective" beliefs (either caused or instantiate by affective states) need not be evaluative: saliently, the belief that one is not thin does not appear to be evaluative, but does often seem to be caused or perhaps instantiated by the affective state of wanting to lose weight. Analogously, in cases of major depressive disorder, the affective state of low self-worth appears to cause or perhaps instantiate non-evaluative beliefs such as "no-one likes me." This suggests that there may be cross-cutting categories, *affective delusions*, and *evaluative delusions* (see also [2, 9]).

[7] For other accounts along these lines, see Craigie and Davies [15]; Heather [38]; Henden [39]; Mele [40, 41].

[8] It is worth stressing that my proposal does not fall down to Pugh's objection. Sue might be unable to ignore the 90% mortality rate, but there is no reason to believe that she is not *capable* of nevertheless deciding to undergo the surgery. The fact that she would not plausibly decide to do so does not entail that she *could* not decide to do so.

[9] It is also worth stressing that, even when they do not, previously expressed preferences are the most significant factor in best interest decisions according to the UK legislation. While doctors tend to treat "best interest" as reducible to "most health-conducive," this is against the spirit and the letter of the legislation—the lifelong values of an incapacitous individual should be respected even when they are "alien" to the clinician.

References

1. Mental Capacity Act, "London: The Stationery Office," 2005.

2. J. Hawkins and L. C. Charland, "Decision-Making Capacity," in *The Stanford Encyclopedia of Philosophy (Fall 2020)*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2020), https://plato.stanford.edu/archives/fall2020/entries/decision-capacity/.

3. G. Meynen, "Depression, Possibilities, and Competence: A Phenomenological Perspective," *Theoretical Medicine and Bioethics* 32 (2011): 181–193.

4. M. Ratcliffe, *Experiences of Depression: A Study in Phenomenology* (OUP Oxford, 2014).

5. M. Parrott, "Bayesian Models, Delusional Beliefs, and Epistemic Possibilities," *British Journal for the Philosophy of Science* 67 (2016): 271–296.

6. M. Parrott, "Delusional Predictions and Explanations," *British Journal for the Philosophy of Science* 72 (2021): 325–353.

7. A. Evans, "Anorexia Nervosa: Illusion in the Sense of Agency," *Mind & Language* 38, no. 2 (2023): 480–494.

8. T. Hope, J. Tan, A. Stewart, and R. Fitzpatrick, "Anorexia Nervosa and the Language of Authenticity," *Hastings Center Report* 41, no. 6 (2011): 19–29.

9. T. Hope, J. Tan, A. Stewart, and J. McMillan, "Agency, Ambivalence and Authenticity: The Many Ways in Which Anorexia Nervosa Can Affect autonomy1," *International Journal of Law in Context* 9, no. 1 (2013): 20–36.

10. L. Bortolotti, *The Epistemic Innocence of Irrational Beliefs* (Oxford University Press, 2020).

11. C. Kong, *Mental Capacity in Relationship: Decision-Making, Dialogue, and Autonomy*, vol. 34 (Cambridge University Press, 2017).

12. J. H. Radden, "Food Refusal, Anorexia and Soft Paternalism: What's at Stake?," *Philosophy, Psychiatry, and Psychology* 28, no. 2 (2021): 141–150.

13. S. Boyle, "How Should the Law Determine Capacity to Refuse Treatment for Anorexia?," *International Journal of Law and Psychiatry* 64 (2019): 250–259, https://doi.org/10.1016/j.ijlp.2019.05.001.

14. W. Aves, "If You Are Not a Patient They Like, Then You Have Capacity": Exploring Mental Health Patient and Survivor Experiences of Being Told "You Have the Capacity to End Your Life," 2022 Psychiatry Is Driving Me Mad.

15. J. Craigie and A. Davies, "Problems of Control: Alcohol Dependence, Anorexia Nervosa, and the Flexible Interpretation of Mental Incapacity Tests," *Medical Law Review* 27, no. 2 (2019): 215–241.

16. L. Smith, "Race, Culture and Capacity—In Conversation With Dr Lade Smith CBE (A. Ruck Keene, Interviewer) [Video]," 2020, https://www.mentalcapacitylawandpolicy.org.uk/race-culture-and-capacity-in-conversation-with-dr-lade-smith-cbe/.

17. J. Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (Oxford University Press, 2020).

18. H. G. Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no. 1 (1971): 5.

19. J. Elster, *Sour Changes: Studies in the Subversion of Rationality* (Cambridge University Press, 1983).

20. A. R. Mele, *Autonomous Agents: From Self-Control to Autonomy* (Oxford University Press, 1995).

21. J. Christman, "Autonomy and Personal History," *Canadian Journal of Philosophy* 21, no. 1 (1991): 1–24.

22. J. Christman, *The Politics of Persons: Individual Autonomy and Socio-Historical Selves* (Cambridge University Press, 2009).

23. G. Dworkin, "The Concept of Autonomy," *Grazer Philosophische Studien* 12, no. 1 (1981): 203–213.

24. S. Y. Kim, "7a. The Place of Ability to Value in the Evaluation of Decision-Making Capacity," in *Philosophy and Psychiatry* (Routledge, 2015), 205–219.

25. J. Tan, A. Stewart, R. Fitzpatrick, and R. Hope, "Competence to Make Treatment Decisions in Anorexia Nervosa: Thinking Processes and Values," *Philosophy, Psychiatry, and Psychology* 13, no. 4 (2006): 267–282.

26. L. C. Charland, T. Hope, A. Stewart, and J. Tan, "Anorexia Nervosa as a Passion," *Philosophy, Psychiatry, and Psychology* 20, no. 4 (2013): 353–365.

27. L. C. Charland, "Passion and Decision-Making Capacity in Anorexia Nervosa," *AJOB Neuroscience* 6, no. 4 (2015): 66–68.

28. J. O. Tan, T. Hope, and A. Stewart, "Anorexia Nervosa and Personal Identity: The Accounts of Patients and Their Parents," *International Journal of Law and Psychiatry* 26, no. 5 (2003): 533–548.

29. O. Abbott, "The Self as the Locus of Morality: A Comparison Between Charles Taylor and George Herbert Mead's Theories of the Moral Constitution of the Self," *Journal for the Theory of Social Behaviour* 50, no. 4 (2020): 516–533.

30. J. Gough, "The Embodied, Relational Self: Extending or Rejecting the Mind?," *Inquiry* 68 (2022): 1–33.

31. G. H. Mead, *Mind, Self and Society* (University of Chicago Press, 1934).

32. K. Wallace, *The Network Self: Relation, Process, and Personal Identity* (Taylor & Francis, 2019), https://books.google.co.uk/books?id=ekmMDwAAQBAJ.

33. A. C. Westlund, "Rethinking Relational Autonomy," *Hypatia* 24, no. 4 (2009): 26–49.

34. J. Christman, "Liberalism, Autonomy, and Self-Transformation," *Social Theory and Practice* 27, no. 2 (2001): 185–206.

35. C. Flores, "Delusional Evidence-Responsiveness," *Synthese* 199, no. 3 (2021): 6299–6330.

36. K. Fulford, "Evaluative Delusions: Their Significance for Philosophy and Psychiatry," *British Journal of Psychiatry* 159, no. S14 (1991): 108–112.

37. L. Bortolotti, *Delusions and Other Irrational Beliefs* (Oxford University Press, 2009).

38. N. Heather, "133Addiction as a Form of Akrasia," in *Addiction and Choice: Rethinking the Relationship*, ed. N. Heather and G. Segal (Oxford University Press, 2016), https://doi.org/10.1093/acprof:oso/9780198727224.003.0008.

39. E. Henden, "Addiction and Autonomy: Why Emotional Dysregulation in Addiction Impairs Autonomy and Why It Matters," *Frontiers in Psychology* 14 (2023): 1081810.

40. A. R. Mele, "Addiction and Self-Control," *Behavior and Philosophy* 24, no. 2 (1996): 99–117.

41. A. R. Mele, "Irresistible Desires," *Noûs* 24, no. 3 (1990): 455.

42. T. M. Breden and J. Vollmann, "The Cognitive Based Approach of Capacity Assessment in Psychiatry: A Philosophical Critique of the MacCAT-T," *Health Care Analysis* 12, no. 4 (2004): 273–283, https://doi.org/10.1007/s10728-004-6635-x.

43. J. Halpern, "Emotions, Autonomy, and Decision-Making Capacity," *AJOB Neuroscience* 2, no. 3 (2011): 62–63.

44. J. Halpern, "When Concretized Emotion-Belief Complexes Derail Decision-Making Capacity," *Bioethics* 26, no. 2 (2012): 108–116.

45. H. Hermann, M. Trachsel, B. S. Elger, and N. Biller-Andorno, "Emotion and Value in the Evaluation of Medical Decision-Making Capacity: A Narrative Review of Arguments," *Frontiers in Psychology* 7 (2016): 765.

46. H. Pickard, "Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder," *Philosophy, Psychiatry, and Psychology* 18 (2011): 209–224.

47. H. Pickard, "Responsibility Without Blame: Philosophical Reflections on Clinical Practice," in *Oxford Handbook of Philosophy and Psychiatry*, ed. M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini, and T. Thorton (Oxford University Press, 2013), 1134–1152.

48. H. Pickard, "Psychopathology and the Ability to Do Otherwise," *Philosophy and Phenomenological Research* 90, no. 1 (2015): 135–163, https://doi.org/10.1111/phpr.12025.

49. H. Pickard, "Responsibility Without Blame for Addiction," *Neuroethics* 10, no. 1 (2017): 169–180.

50. T. Minkowitz, "CRPD Article 12 and the Alternative to Functional Capacity: Preliminary Thoughts Towards Transformation," *SSRN Electronic Journal* (2013).

51. A. Ruck Keene, N. B. Kane, S. Y. H. Kim, and G. S. Owen, "Mental Capacity—Why Look for a Paradigm Shift?," *Medical Law Review* 31, no. 3 (2023): 340–357, https://doi.org/10.1093/medlaw/fwac052.

52. Great Britain: Department for Constitutional Affairs, *Mental Capacity Act 2005 Code of Practice: [Large Print 2007 Final Edition]* (Stationery Office, 2007), https://books.google.co.uk/books?id=imknbETujCIC.

53. S. Boyle, A. McGee, and F. Wood, "How to Determine the Capacity of a Person With Depression Who Requests Voluntary Assisted Dying," *Psychiatry, Psychology and Law* 32, no. 2 (2025): 276–290, https://doi.org/10.1080/13218719.2023.2296486.

54. G. S. Owen, F. Freyenhagen, M. Hotopf, and W. Martin, "Temporal Inabilities and Decision-Making Capacity in Depression," *Phenomenology and the Cognitive Sciences* 14, no. 1 (2015): 163–182.

55. G. Owen, A. S. David, G. Richardson, G. Szmukler, P. Hayward, and M. Hotopf, "Mental Capacity, Diagnosis and Insight in Psychiatric In-Patients: A Cross-Sectional Study," *Psychological Medicine* 39, no. 8 (2009): 1389–1398, https://doi.org/10.1017/S0033291708004637.

56. W. Martin and S. Gurbai, "Surveying the Geneva Impasse: Coercive Care and Human Rights," *International Journal of Law and Psychiatry* 64 (2019): 117–128, https://doi.org/10.1016/j.ijlp.2019.03.001.

57. J. Fiala-Butora, "The Influence of the Convention on the Rights of Persons With Disabilities on the European Court of Human Rights in the Area of Mental Health Law: Divergence and Unexplored Potential," *International Journal of Law and Psychiatry* 94 (2024): 101965, https://doi.org/10.1016/j.ijlp.2024.101965.

58. G. Quinn, "Personhood & Legal Capacity: Perspectives on the Paradigm Shift of Article 12 CRPD," 20 (2010): 3–5.

59. P. Gooding, "Navigating the 'Flashing Amber Lights' of the Right to Legal Capacity in the United Nations Convention on the Rights of Persons With Disabilities: Responding to Major Concerns," *Human Rights Law Review* 15, no. 1 (2015): 45–71, https://doi.org/10.1093/hrlr/ngu045.

60. L. Series and A. Nilsson, "Chapter Article 12 CRPD: Equal Recognition Before the Law," in *The UN Convention on the Rights of Persons With Disabilities* (Oxford University Press, 2018).

61. P. Skowron, "Giving Substance to 'the Best Interpretation of Will and Preferences'," *International Journal of Law and Psychiatry* 62 (2019): 125–134, https://doi.org/10.1016/j.ijlp.2018.12.001.

62. J. Ahlin, "The Impossibility of Reliably Determining the Authenticity of Desires: Implications for Informed Consent," *Medicine, Health Care and Philosophy* 21, no. 1 (2018): 43–50.

63. A. L. Burton, R. Brown, and M. J. Abbott, "Overcoming Difficulties in Measuring Emotional Regulation: Assessing and Comparing the Psychometric Properties of the DERS Long and Short Forms," *Cogent Psychology* 9, no. 1 (2022): 2060629, https://doi.org/10.1080/23311908.2022.2060629.

64. S. Paradiso, J. N. Beadle, V. Raymont, and J. Grafman, "Suicidal Thoughts and Emotion Competence," *Journal of Clinical and Experimental Neuropsychology* 38, no. 8 (2016): 887–899, https://doi.org/10.1080/13803395.2016.1172558.

65. S. E. George, A. C. Page, G. R. Hooke, and W. G. Stritzke, "Multifacet Assessment of Capability for Suicide: Development and Prospective Validation of the Acquired Capability With Rehearsal for Suicide Scale," *Psychological Assessment* 28, no. 11 (2016): 1452–1464.

66. T. Joiner, *Why People Die by Suicide* (Harvard University Press, 2005).

67. E. A. Selby, M. D. Anestis, T. W. Bender, et al., "Overcoming the Fear of Lethal Injury: Evaluating Suicidal Behavior in the Military Through the Lens of the Interpersonal–Psychological Theory of Suicide," *Clinical Psychology Review* 30, no. 3 (2010): 298–307.

68. P. N. Smith and K. C. Cukrowicz, "Capable of Suicide: A Functional Model of the Acquired Capability Component of the Interpersonal-Psychological Theory of Suicide," *Suicide and Life-Threatening Behavior* 40, no. 3 (2010): 266–275.