

Binding the Self:
The Ethics of Ulysses Contracts

Andrew Franklin-Hall

The University of Toronto

andrew.franklin.hall@utoronto

Penultimate Draft

“Binding the Self: The Ethics of Ulysses Contracts,” *Ethics* 134:1 (October 2023), 57-76.

1. Introduction

We frequently fear that our present intentions will be defeated by our own future actions. You intend to abstain from that second drink but know that, once the first glass is drained, the second will look quite tempting. What can we do about this? Well, when we don't fully trust someone else to do as they say, we may seek to bind them by contract. Similarly, when we lack faith in ourselves, we might try self-binding. But, as it turns out, it can be hard to tie a knot you cannot loosen. In one of the charming *Frog and Toad* children's stories, the two amphibious friends decide they've had enough cookies for one day yet find they can't stop eating them. Frog has the idea of putting the cookies in a box and tying it up with string. “But,” Toad objects, “we can cut the string and open the box.”¹ When you find yourself in a predicament of this sort, you might want someone else to tie your hands for you—as Ulysses did. Warned that the Sirens' call will lure his sailors overboard to their doom, Ulysses orders his crew to plug their ears with wax. However, wanting to hear the enchanting song for himself, Ulysses orders his men to bind him

¹ Arnold Lobel, *Frog and Toad Together* (New York: Harper and Rowe, 1971).

to the mast and ignore his later pleas for release until they are out of earshot of the Sirens.² In reference to this emblematic story, we can bestow the name of “Ulysses contract” on any arrangement in which (roughly) A gets B to agree to intervene at a later time in order to force or pressure A to abide by his former aims.

Prima facie, the very idea of a Ulysses contract is puzzling. The general assumption seems to be that, absent A’s authorization at time t_1 , it would be wrong for B to force or pressure A, or act for A, at t_2 . But if A’s will is so important, why should A’s will at t_2 be ignored? True, A at t_1 may tell B to disregard what he says at t_2 . But he’ll do just the reverse at t_2 . What is so special about A’s will at t_1 ?

On the view I endorse, Ulysses contracts are valid if and only if A’s will at t_1 is a better expression of the values and aims we can continue to attribute to him at t_2 than is his actual will at t_2 . Call this the “Authentic Authorization Model.” Essentially, a person can authorize someone to bind his future self against various forms of backsliding and error, but not against acting contrary to his prior intentions once he has genuinely changed his mind about what he is committed to. I trust this is an intuitive view. And yet, for two reasons, it has not received a satisfactory philosophical treatment.

The first is that writers have struggled to fit together two conceptions of autonomy at work on this picture. On the one hand, it might seem like the main rationale for Ulysses contracts derives from the value of “authenticity”: that of protecting a person from acting contrary to his own settled values and aims. On the other hand, the idea that Ulysses has a discretionary power to authorize others to bind him to the mast is rooted in a conception of authority or

² Homer, *The Odyssey*, Book XII.

“sovereignty”: others have permission to intervene in Ulysses’ sphere of control only because he has given them permission to do so.³ Yet many writers have tried to make sense of Ulysses contracts by appealing to just one of these conceptions of autonomy. Such views have the merit of simplicity, but (as I will show) they have undesirable implications. A more adequate view, as I shall argue, will incorporate the values of sovereignty and authenticity.

The other stumbling block concerns the difficulty of articulating the distinction between backsliding and changing one’s mind. Some writers in this literature have suggested that the key to the distinction is that, unlike backsliding, changing one’s mind is a rational process. I will argue that, in routine cases, the essential question is whether a new configuration of values and aims has taken firm root, whether rationally or not. To count as a genuine change of mind, though, a person must be responsible, or “answerable” (as I will say), for the fact that she no longer endorses her former values—a condition that may not hold when we consider people who have suffered serious psychological impairments.

I proceed as follows. After clarifying the concept of the Ulysses contract and staking out some of my assumptions (§2), I outline four simple views about Ulysses contracts we find in the literature and explain why none of them is acceptable (§3). Next, I set out the approach I favor. First, I offer an account of the proper relation between sovereignty and authenticity in Ulysses contracts (§4). I then make the case that we ought to conceive of a person as having changed her mind in a way sufficient to free her from her Ulysses contract if and only if her formerly authoritative set of values and concerns has been stably replaced by a new set (§5) and the

³ Two recent papers usefully delineate what I call “sovereignty” and “authenticity”: Daniel Brudney and John Lantos, “Agency and Authenticity: Which Value Grounds Patient Choice,” *Theoretical Medicine and Bioethics* 32 (2011), 217-227; and David Enoch, “Hypothetical Consent and the Value(s) of Autonomy,” *Ethics* 128 (2017), 6-36.

subject is answerable for having abandoned those former values and concerns (§6). There follows a brief conclusion (§7).

2. Preliminaries

There are different ways and reasons to tie one's hands. My purpose in this section is to clarify the forms of self-binding that concern us and to state a few assumptions.

The term "Ulysses contract" was introduced to refer to so-called "voluntary commitment contracts" or "psychiatric advance directives": legal devices permitting patients with episodic mental illnesses to bind themselves in advance to hospitalization and treatment in case of relapse, ignoring their later protests.⁴ Similar instruments have been proposed for managing self-control problems like addiction.⁵ However, adhering to recent usage, I am going to understand the term to designate any arrangement with the same basic structure, regardless of subject matter and legal enforceability. In taking this broad view, I follow writers like Allen Buchanan and Dan Brock who observe that psychiatric advance directives are really just

legal devices to carry out what people more commonly do in informal, everyday contexts in which they anticipate impaired judgment. For example, ... a wife dieting may put

⁴ The idea is raised in Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality*, revised edition (Cambridge: Cambridge University Press [1979], 1984), and Charles M. Culver and Bernard Gert briefly refer to an "Odysseus Pact" in *Philosophy in Medicine: Conceptual and Ethical issues in Medicine and Psychiatry* (New York: Oxford University Press, 1982), 174. Timothy Howell, Ronald J. Diamond, and Daniel Wikler develop the proposal in "Is There a Case for Voluntary Commitment," in Tom Beauchamp and LeRoy Walter (eds.), *Contemporary Issues in Bioethics*, 2nd Edition (Belmont, CA: Wadsworth Publishing Company, 1982). For critical discussion, see Rebecca Dresser, "Ulysses and the Psychiatrists: A Legal and Policy Analysis of the Voluntary Commitment Contract," *Harvard Civil Rights-Civil Liberties Review* 16 (1981-1982), 777-854, and Dresser, "Bound to Treatment: The Ulysses Contract," *Hastings Center Report* 14 (1984), 13-16.

⁵ Thomas C. Schelling, "The Intimate Contest for Self-Command" and "Ethics, Law, and the Exercise of Self-Command," both in his *Choice and Consequence: Perspectives of an Errant Economist* (Cambridge, MA: Harvard University Press, 1984). See also Jon Elster, *Ulysses Unbound* (Cambridge: Cambridge University Press, 2000), 66-68 and Chrisoula Andreou, "Making a Clean Break: Addiction and Ulysses Contracts," *Bioethics* 22 (2008), 25-31.

particularly tempting and fattening foods in the hands of her husband with instructions not to serve her more than a specified amount even if she later asks for more.⁶

Although much of the literature on Ulysses contracts is focused on the psychiatric context, I will consider the matter from a more general vantage point since it raises interesting questions about autonomy. But this means I largely sidestep practical and legal questions relevant to mental health policy.

A Ulysses contract is a kind of “precommitment,” which Jon Elster characterizes as the act of making a decision at time t_1 that sets into motion some causal process that has the effect of removing certain options or making them costlier, with the purpose of increasing the likelihood that one will carry out another decision at t_2 , which one could have made anyway, but may not have.⁷ To this, we should add that sometimes the point of precommitment is to lock in a certain course of action, not to influence one’s future *decision*. The purpose of the voluntary commitment contract, after all, is to take the future choice out of my hands entirely.⁸

Precommitment works by manipulating, coercing, or boxing-in the future self. Such “self-enforcement” differs from “self-command”—where one exercises self-control agentially by mastering errant inclinations through decision, attention, or will-power.⁹ Thus, some recovering alcoholics maintain control chiefly by keeping their resolution continually in mind (perhaps by carrying a special token). That is not precommitment. Taking the drug disulfiram, on the other

⁶ Allen E. Buchanan and Dan W. Brock, *Deciding for Others* (Cambridge: Cambridge University Press, 1989), 357.

⁷ Elster, *Ulysses and the Sirens*, 37-47. I omit some subtleties irrelevant to my discussion.

⁸ Dan W. Brock, “Precommitment in Bioethics: Some Theoretical Issues,” *Texas Law Review* 81 (2002-2003), 1805-1821.

⁹ See Theo van Willigenburg and Patrick J.J. Delaere, “Protecting Autonomy as Authenticity Using Ulysses Contracts,” *Journal of Medicine and Philosophy* 30 (2005), 395-409.

hand, which makes the subject nauseous if he consumes alcohol, *is* precommitment, since it imposes a sanction on the undesired action.

Two things separate Ulysses contracts from other forms of precommitment. First, they involve *self-management*: their purpose is to help you follow through on certain aims or resolutions you fear you might abandon. This contrasts with *strategic precommitment*, which aims to force an adversary's hand by taking some of your own options off the table.¹⁰ Second, unlike the *unilateral precommitment* of the disulfiram case, Ulysses contracts constitute *cooperative precommitment*: they authorize another party to intervene on the principal's behalf in some way that would otherwise be impermissible or inappropriate. Thus, Ulysses precommitted, not by getting himself tangled up in the rigging, but by ordering his men to bind him to the mast. Derek Parfit's case of the Russian Nobleman is another example.¹¹ A young aristocrat expects to inherit the family estate one day and, inspired by socialist ideals, wants to give it away to the peasants. However, he fears that, when the inheritance finally comes, he will have grown as conservative as the other middle-aged men of his class. So, having apparently retained the legal counsel of Rube Goldberg, the young nobleman signs a document that will automatically transfer the estate to the peasants upon his receipt of it, and which only his wife can revoke. He then gets his wife to promise him she won't nullify the deed, even if he should later implore her to do so. Naturally, unilateral precommitment and Ulysses contracts raise some similar questions, but my focus will be on the latter, partly to avoid complexities about duties to

¹⁰ Jon Elster, "Don't Burn Your Bridge Before You Come to It," *Texas Law Review* 81 (2002-2003), 1761-1765; Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960).

¹¹ Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, [1984] 1986), 327.

self, but mainly because cooperative self-binding brings questions about authority into sharper focus.

Now, a couple of clarifications. Asking someone to remind you of your resolution in a moment of temptation is not a Ulysses contract, even if, later, you don't welcome the reminder—*unless* the point is to impose a mild penalty on yourself. Ulysses contracts always involve some sort of coercion, sanction, compulsion, or preemption of choice. Likewise, to simply involve someone in your scheme of self-maintenance does not make it a Ulysses contract. Frog and Toad finally avoid eating the cookies by feeding them to the birds, and you can avoid spending your money by loaning it out. In a Ulysses contract, however, the other party must *aim* to assist Ulysses in self-management.

What exactly is “contractual” about Ulysses contracts? That language is rooted in the idea of a legally enforceable document to bind oneself to treatment. But this can be misleading. Standard contracts are agreements for mutual benefit. Psychiatric advance directives are more like wills, particularly “living wills”: they instruct others to carry out a person's instructions on his behalf. Talk of a “contract” may nonetheless seem fitting, since a person making a Ulysses contract limits his future freedom. But it's best to just treat “Ulysses contract” as a term of art and specify its structure afresh.

The standard Ulysses contract involves the exercise of several simple normative powers.¹² First, the principal authorizes another party—the *deputy*—to restrain the principal in some way that would ordinarily be wrong or inappropriate (given the nature of their relationship). Second,

¹² Normative powers—like consenting, promising, transferring, etc.—enable us to change, by declaration, certain permissions, requirements, and prohibitions. See Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1975).

the principal waives his future ability to rescind that authorization, at least in a range of circumstances. Third, the principal sometimes secures the deputy's obligation to follow the principal's instructions. Finally, although normative powers are usually discussed in interpersonal terms, you could think I change my own normative situation when I make a resolution or vow, creating a new duty for myself.¹³ Ulysses contracts often involve such resolutions, but they need not if the point is to have someone prevent you from doing something or make a choice for you.

Notice that, when we ask whether Ulysses contracts ought to be enforced, any number of considerations may be relevant. If Clyde, afraid of losing his nerve, asks Bonnie to coerce him into carrying out his planned bank heist, she ought to refuse partly out of concern for others. Or, if Ron wants euthanasia once he has advanced dementia, even should he then want to live, it may be asking too much of his family to carry out his will. My narrower focus, however, will just be on how the principal's autonomy bears on the question of Ulysses contracts.

Lastly, I presume throughout that we are dealing with one person persisting over time. Some philosophers hold that, if someone's psychology changes enough, then the earlier and the later selves are, metaphysically, two distinct persons. In that case, the earlier self would apparently have no authority over the later self.^{14, 15} But it isn't plausible that the changes in perspective we encounter in most Ulysses cases sever personal identity. For, even accepting the debatable

¹³ I remain neutral concerning the recent debate about whether all future-oriented intentions create new reasons.

¹⁴ Rebecca Dresser makes this argument regarding advance directives for permanently incompetent patients, but not in her essays on psychiatric advance directives. For the former, see her "Life, Death, and Incompetent Patients: Conceptual Infirmities and Hidden Values in the Law," *Arizona Law Review* 28 (1986), 373-405, and "Dworkin on Dementia: Elegant Theory, Questionable Policy," *Hastings Center Report* 25, 32-38.

¹⁵ The Russian Nobleman case might seem to suggest otherwise, but Parfit appeals to personal identity to explain why the Nobleman cannot release his wife from his earlier instructions, not to explain why it is legitimate for him to prospectively give away the estate. Parfit agrees that my promise would not bind "my" numerically distinct, future self (*Reasons and Persons*, 327).

“psychological theory” of personal identity underlying this argument, we must ask how significant the psychological transformation must be to constitute a veritable break in the continuity of the person.^{16, 17} Cases like advanced dementia, involving radical memory loss and substantial changes in character, *might* meet that threshold, but most Ulysses cases don’t approach that kind of psychological rupture.

3. Four Simple Models of Ulysses Contracts

As I see it, most existing accounts of Ulysses contracts may be understood as versions of the following four views: the Sovereignty Model, the Competence Model, the Diachronic Autonomy Model, and the Authenticity Model. These approaches are all “simple” in that they each work with a single notion of autonomy. Having them on the table, along with some key problems with them, will help me articulate and motivate the more complex view I defend.

A. The Sovereignty Model

Typically, you can rescind consent and revoke commands at any time. This ability is axiomatic to our current understanding of sexual and medical consent. Ulysses contracts are, therefore, puzzling. How can a person be capable of authorizing someone at one moment and lack the ability to countermand the authorization later?

Well, you might think the ability to bind your future self simply follows from your authority over your own life. As Schelling says, “Full freedom entails the freedom to bind oneself, to incur

¹⁶ Alternatives to the psychological theory include the embodied mind and animalist accounts. See Jeff McMahan, *The Ethics of Killing: Problems at the Margins of Life* (New York: Oxford University Press, 2000), ch. 1, and David DeGrazia, *Human Identity and Bioethics* (Cambridge: Cambridge University Press, 2005), chs. 2 & 5.

¹⁷ On the threshold question, see Buchanan and Brock, *Deciding for Others*, ch. 3.

obligation, to reduce one's range of choice."¹⁸ The most important champion of this *Sovereignty Model* may be Joel Feinberg. Feinberg conceives of personal sovereignty as the right to make choices about your own body, life, and possessions (without doing harm to others), and this includes the ability to forfeit certain rights and powers.¹⁹ So, just as your personal sovereignty enables you to renounce some of your rights in a bilateral contract, so too you can authorize me to bind you to your current will and forfeit your right to withdraw the authorization.

When the earlier self in a fully voluntary way ... explicitly instructs another, as in the Odyssean example, not to accept contrary instructions from the future self, then the earlier choice, being the genuine choice of a sovereign being, free to dispose of his own lot in the future, must continue to govern.²⁰

Ulysses contracts are binding, then, not only when a person wants to restrain his future self from (say) weakness of will, but even when his future self would be acting in a subjectively rational and voluntary way. In fact, Feinberg explicitly holds that the young Russian Nobleman should be able to bind his future self against a genuine change of heart.²¹ Of course, a person might use this ability to bind himself unwisely, or in a way he will later regret, but the same is true of any contractual agreement. To deny a person this ability on such grounds would be objectionably paternalistic.

How attractive is this picture? Not very. First, I don't think our intuitions about freedom of contract really support it. A standard contract binds me because I am bound to another party. But to whom am I bound in a Ulysses contract? (1) You might think I am bound to the person

¹⁸ Thomas Schelling, "The Exercise of Self-Command," 98.

¹⁹ Joel Feinberg, *Harm to Self* (New York: Oxford University Press, 1986), 54, 68-70.

²⁰ Feinberg, *Harm to Self*, 83.

²¹ Feinberg, *Harm to Self*, 86.

binding me. But this doesn't seem right. My non-compliance with the Ulysses contract does not wrong *him*, and he cannot release me at his pleasure. He is, rather, acting in my name. (2) If I am ultimately bound *to myself* though, why can't I release myself?²² (3) Am I bound *to my past* self? But if I am said to be distinct from my past self, then we seem to be making dubious assumptions about personal identity. (4) So, maybe I can simply relinquish certain rights and powers without binding myself to anyone—they just vanish. But, if that's so, then it seems like Ulysses contracts actually have little in common with ordinary contracts—in which case we can't appeal to the one to support the other.

Moreover, we do not usually suppose people have an unlimited power to forfeit their future ability to change their minds. Of course, where your commitment creates legitimate interests on the part of others, your freedom to withdraw from it may be limited. But when we are dealing with permissions and authorizations, we don't usually assume you can forfeit the right to change course. You cannot give someone standing sexual consent and relinquish any power to rescind it. You cannot write a will that prevents you, though still of sound mind, from changing it. Nor, again, can you give someone power of attorney and forfeit the ability of your competent future self to withdraw it. In fact, it is notable that, even when we *are* dealing with cases where the interests of others *are* concerned, our laws show special solicitude for protecting the ability to reconsider. For example, you generally cannot enter an employment contract you cannot get out of, and marriage law has moved in the same direction with the adoption of no-fault divorce.

²² As Hobbes says, "he is free, that can be free when he will: nor is it possible for any person to be bound to himself; because he that can bind can release; and therefore he that is bound to himself only, is not bound" (*Leviathan*, ch. XXVI.6).

The Sovereignty Model has another counterintuitive implication. What if on Monday, Bud, though competent, is angry and wants Lou to bind his calmer self on Friday to some course of action that conflicts with his settled values? If the power to bind yourself were just like the sovereignty you have over your self-regarding actions, or like the freedom of contract, then, according to the Sovereignty Model, Bud ought to be able to bind his future, more rational self. After all, a competent person is free to act or make agreements unreasonably, even if that does constrain his future options.²³

Feinberg balks at this conclusion, holding that an unreasonable person cannot bind his more reasonable future self, since the more reasonable, later self represents the “real,” “enduring self.” That’s sensible, but this appeal to “authenticity” marks a departure from the logic of the Sovereignty Approach. If authenticity is so important, why does it only show up here? Why allow—as in the Russian Nobleman case—a person to bind a future, equally authentic self who has different settled preferences?

One may also object that the Sovereignty Model relies on a one-sided understanding of autonomy. As Jennifer Radden has argued, autonomy, intuitively, isn’t just about making commitments and sticking to them; it also concerns our “ability to entertain second thoughts, to reconsider, adapt and change direction in the light of a new piece of information, or a telling experience.”²⁴ This is why it would be wrong to force someone to remain faithful to a ten-year-old vow of celibacy once he has come to sincerely and stably disavow it. But Feinberg would say

²³ Feinberg, *Harm to Self*, 109.

²⁴ Jennifer Radden, “Second Thoughts: Revoking Decisions Over One’s Own Future,” *Philosophy and Phenomenological Research* 54 (1994), 787-801, 792.

that, if we were sticklers about autonomy, then we *should* hold him to it.²⁵ That seems preposterous.

B. The Competence Model

According to the *Competence Model*, a competent person has the authority to bind his future self if and only if the future self is incompetent. While you remain competent, you always retain your authority to reconsider your previous decisions. The reason your later self cannot countermand your prior authorization in a valid Ulysses contract is straightforward: you lack the competence to reconsider your former decision and exercise the power. The deputy, on this view, is conceived as the principal's agent tasked with carrying out Ulysses' will when Ulysses cannot do so himself.

The Competence Model is frequently held to govern traditional end-of-life advance directives, and some (including Buchanan and Brock) have extended it to cover incapacitating mental illnesses and impairments.²⁶ If offered as a *general theory* of Ulysses contracts, however, it seems to prohibit more self-binding than seems reasonable. Competence, after all, is not a very demanding standard. As one medical ethics textbook puts it, ““Seemingly competent and autonomous patients can choose irrationally by [for example] thinking about the immediate future and ignoring long-term risks, believing that nothing bad will happen to them, [or] acting on unreasonable fears that make them avoid unnecessary treatment.”²⁷ For competence, I

²⁵ Feinberg says the Russian Nobleman's wife has an obligation to hold her husband to his former aim, allowing only that “Perhaps ... [she] ought to relent, and deliberately decline to do her duty, out of pity, or from simple humanity” (*Harm to Self*, 87).

²⁶ Buchanan and Brock, *Deciding for Others*, 350-357; Brock, “A Proposal for the Use of Advance Directives in the Treatment of Incompetent Mentally Ill Patients,” *Bioethics* 7 (1993), 247-256.

²⁷ Philip C. Hébert and Wayne Rosen, *Doing Right: A Practical Guide to Ethics for Medical Trainees and Physicians, Fourth Edition* (Oxford: Oxford University Press, 2020), 61-62. See also Brock and Steven A. Wartman, “When

assume it is enough that the person has the *ability*, in a given situation, to make an informed and reasoned decision guided by his own values and commitments—whether he actually does so or not. But the kinds of everyday irrationality just mentioned, though they do not undermine competence, are just the sorts of things it seems sensible to want to bind oneself against. If Felix has been drinking more than he likes on weeknights, surely he can ask Oscar, his roommate, to hide his liquor from him until the weekend, even should he later, in a moment of weakness, ask for a drink. Indeed, most of the examples of self-binding explored in the seminal works of Elster and Schelling (e.g., difficulty saving money, problems of emotional control, bad habits) do not involve incompetence.

You might think we could handle this objection by adopting a more demanding standard of competence. In fact, Buchanan and Brock say that, in medicine, a “helpful way of characterizing the proper aim of the evaluator of ... competence ... is to think of him or her addressing the patient in this fashion: ‘...Help me to see whether your choice is reasonable, not in the sense that it is what I or most people would choose, but that it is reasonable for you in light of your underlying and enduring aims and values.’”²⁸ Taken at face value, this suggests that substantial subjective irrationality undermines competence. This certainly would license Ulysses contracts in a wider range of cases. But if paternalistic interference is justified whenever an incompetent agent is harming her own interests, it would also greatly expand the scope of legitimate

Competent Patients Make Irrational Decisions,” in Brock, *Life and Death: Philosophical Essays in Biomedical Ethics* (Cambridge: Cambridge University Press, 1993).

²⁸ Buchanan and Brock, *Deciding for Others*, 56.

paternalism. My sense is that everyday irrationality is a good reason to make a Ulysses contract, but that it does not make you incompetent or liable to unsolicited paternalism.

One could say that the Competence Model is best understood as an account specifically intended to apply to legally enforceable advance directives where strictness is a virtue, since we want safeguards against error and abuse. Since I am after a general theory of Ulysses contracts, I could accept this. But I have two reservations. First, if psychiatric advance directives really are (as Buchanan and Brock say) just special cases of “what people commonly do in informal, everyday contexts,” then you might think the general theory will have some bearing on the relevant legal framework. Second, and more importantly, even in psychiatric cases, the Competence Model may license too little, too late. If I’m worried about what I might do in a manic episode, I may want to be hospitalized as soon as my symptoms appear. These symptoms—sleeping little, speaking rapidly about ambitious plans, etc.—may be harbingers of what is to come without yet rendering me incompetent. If we wait until I really am incompetent, I may have already put myself in some dangerous situations and have done irreversible damage to my finances and relationships. While it is true that I am already in the *process* of becoming incompetent in this case, permitting me to bind myself in such circumstances would be a major departure from the basic logic of the Competence Model, since I am not yet incompetent to exercise the relevant normative powers.

Two other objections should be noted. The first is that the Competence Model is in danger of making Ulysses contracts redundant. Arguably, we already have “soft paternalist” grounds for preventing people from acting contrary to their standing interests when they are incompetent

without needing their authorization.²⁹ A Ulysses contract that licenses intervention only under those same circumstances does not contribute much.³⁰ The second is that, if incompetence is understood to represent a sufficient as well as necessary condition for self-binding, then the Competence Model may give Ulysses too much authority over his future self. The prototype for the Competence Model is the living will, but in those cases the future self (being unconscious or nearly so) has no evaluative perspective to speak of. In the typical Ulysses contract, however, even if the future self lacks decision-making capacity, he may still have certain values or concerns of his own, and the aims of his past self might seem quite alien to him. It seems problematic, then, if Ulysses is not required to take that future evaluative perspective into account at all.³¹ This last point might lead you to the view that, if the later self is competent, he can always countermand his prior instructions, while if he is incompetent, we should just attend to his welfare interests going forward—never mind what he used to want.³² Before we accept that conclusion, however, let us see if we can find a more convincing account of Ulysses contracts.

²⁹ “The harm to others principle permits us to protect a person from the choices of other people; soft paternalism would permit us to protect him from ‘nonvoluntary choices,’ which, being the genuine choices of no one at all, are no less foreign to him” (Feinberg, *Harm to Self*, 12).

³⁰ In most jurisdictions, however, a person may be hospitalized and treated for mental illness against his will only if he poses an immediate danger to his own life or to the safety of others. In that context, a Competence-Model Ulysses contract would make a difference. See Buchanan and Brock, *Deciding for Others*, 317-325.

³¹ The Competence Model would also (like the Sovereignty Model) seem to permit a competent agent to bind his future incompetent self to an irrational course of action. An amendment might be proposed, restricting the authority to competent-and-reasonable agents. But this, again, introduces elements foreign to the basic logic of the Competence Model. Furthermore, it does not solve the other problems raised above.

³² See Dresser, “Life, Death, and Incompetent Patients”; and Dresser and John A. Robertson, “Quality of Life and Non-Treatment Decisions for Incompetent Patients: A Critique of the Orthodox Approach,” *Law, Medicine, and Health Care* 17 (1989), 234-244.

C. The Diachronic Autonomy and Authenticity Models

The kind of autonomy that warrants respect on the Sovereignty and Competence Models is content-independent: what matters is just that the agent has the requisite abilities to exercise his agency. But you could also say that a choice counts as autonomous only if its content suitably expresses, or at least is consistent with, the agent's stronger preferences or deeper values.

Consider, first, the *Diachronic Autonomy Model*. John Davis argues that our true interest in self-determination lies, not in being able to act autonomously at a particular time, but in having our circumstances conform to what we concurrently most want *over time*.³³ Suppose Juan gets a tattoo because that is what he most wants. Synchronically, the decision was perfectly autonomous. But, soon thereafter, he comes to stably regret it. Davis thinks that, considered diachronically, Juan's choice injured his interest in autonomy. This leads Davis to endorse what he calls the "Diachronic Respect Principle": in responding to a person's interest in autonomy, we should "act so that the circumstances of the agent's life are consistent with his or her concurrent wishes over the longest period of time, qualified by reference to the intensity of the wishes and the severity of the conflict."³⁴

Now, suppose Ulysses has certain standing preferences, but because he also undergoes temporary preference reversals due to mental illness or ordinary temptation, he wants to bind himself against acting on those short-lived preferences which work against his more abiding

³³ John K. Davis, "How to Justify Enforcing a Ulysses Contract When Ulysses is Competent to Refuse," *Kennedy Institute of Ethics Journal* 18 (2008), 87-106, 88-89.

³⁴ Davis, "Enforcing a Ulysses Contract," 98. By "concurrent wishes," Davis means desires satisfied while the desire is still held.

aims.³⁵ According to Davis, we are justified in helping him, even if he remains competent, since that would maximize the satisfaction of his concurrent strongest preferences over time. But we shouldn't help someone bind himself, even if the later self is incompetent, if this won't help him get what he, going forward, will concurrently most want.^{36,37}

The *Authenticity Model*, by contrast, takes a synchronic perspective. This is a picture defended in a couple of papers by Julian Savulescu, one co-authored with Donna Dickenson.³⁸ For Savulescu, it is objectionably paternalistic to frustrate a person's present preferences in the name of his future preferences; respect for autonomy should always be oriented to a person's present preferences.³⁹ But he thinks, quite plausibly, that we should conceive of present preferences dispositionally. Just as a person continues to have preferences when asleep, he continues to have certain preferences when they are suppressed by other occurrent desires. Of course, we have many dispositional preferences in this sense. The ones warranting our respect are those that are most authentically the agent's own. These represent what a person truly values or cares about, rather than what he merely desires. For Savulescu, our values are equivalent to what we "rationally desire": what we prefer after having vividly imagined the alternatives with

³⁵ Cf. George Ainslie, *Breakdown of Will* (Cambridge: Cambridge University Press, 2001).

³⁶ Relatedly, Andreou ("Making a Clean Break") argues that, in light of certain patterns of preference reversal or intransitivity, submitting to constraint now can be justified in terms of Ulysses' current preferences for the future.

³⁷ The Diachronic Autonomy Model has a decidedly consequentialist cast. Davis insists that the theory is not about making Ulysses "better off, or even happier, but just to help bring about circumstances that are closer to the way he will want them to be over time" ("Enforcing a Ulysses Contract," 102). But this, I simply note, is just what it means to make someone better off on "concurrent" desire-fulfillment theories.

³⁸ Julian Savulescu, "Rational Desires and the Limitation of Life-Sustaining Treatment," *Bioethics* 8 (1994), 191-222; Savulescu and Donna Dickenson, "The Time-Frame of Preferences, Dispositions, and the Validity of Advance Directives for the Mentally Ill," *Philosophy, Psychology, and Psychiatry* 5 (1998), 225-246.

³⁹ Savulescu and Dickenson, "The Time Frame of Preferences," 230.

all the available, relevant information in hand and without having made any logical errors.⁴⁰ But the basic model is consistent with other conceptions, so long as we can distinguish a person's occurrent desires from the values representing the more authentic and authoritative standpoint of the agent.

The application of the Authenticity Model to Ulysses contracts is straightforward. We ought to help a person bind his future self whenever doing so helps him remain true to values he continues to hold at that future time. In other words, we should help people avoid backsliding or otherwise straying from their values. But we should not hold a person to a Ulysses contract if he no longer has the values (in a dispositional sense) that motivated the Ulysses contract in the first place.

Why respect a person's *present values*? If the fulfillment of someone's concurrent values is related to her well-being, then we ought to care equally about her values at all points in time going forward—as the Diachronic Autonomy Model would have it. On the Authenticity Model, we are supposed to respect the person as she is now. But this kind of respect differs from respect for a person's agency or sovereignty. *Sovereignty respect* defers to the actual choices of a competent agent, whether these reflect her deeper values or not. What we are talking about, here, is not respect for a person's decision-making authority, but rather respect for the values that the person regards as possessing authority to guide her decisions. Call that *authenticity respect*. This might seem like an idea awkwardly sandwiched between diachronic autonomy and sovereignty respect. But the intuition, I take it, is that a person's values represent his authentic

⁴⁰ Savulescu, "Rational Desires."

standpoint as an agent, and it is possession of this standpoint that makes someone a “person,” a distinctively human agent, capable of self-governance. Therefore, to respect a person’s present values is to side with this “higher,” more authoritative self that makes the actor a person in the full sense.

The idea that self-binding has something to do with locking in our more authentic preferences is quite common in the Ulysses contract literature. Still, some will regard it as an advantage of the Diachronic Autonomy Model that it does not need to appeal to authenticity—an idea some find dubious. They prefer the more hard-headed approach of simply observing the changes in people’s expressed preferences over time. But, from the first-person perspective, the idea that some of my preferences are deeper and have more authority in my deliberation seems quite natural. This is not to deny that sometimes our preferences simply fluctuate, but this feels very different than weakness of will or otherwise being out of touch with what we really care about.⁴¹

In any case, the Diachronic Autonomy and Authenticity Models typically reach the same conclusions about cases. They diverge only when: (i) a person’s less authentic preferences are activated more often than his authentic ones; or (ii) we know that a person has a certain preference now, will have it again in the future, but will not have it, even dispositionally, in the interim when we must intervene. Both approaches, in one way, grant Ulysses *more* authority over his future self than the Competence Model, since they can justify binding a competent

⁴¹ See for instance: Harry Frankfurt, “Freedom of the Will and the Concept of a Person,” *Journal of Philosophy* 68 (1971), 5-20; Gary Watson, “Free Agency,” *Journal of Philosophy* 72 (1975), 205-220; Michael Bratman, “Reflection, Planning, and Temporally Extended Agency,” *The Philosophical Review* (2000), 35-61.

future self. In another way, however, both approaches rein in the principal's authority, since they deny him the power to impose preferences on his later self that he will no longer hold.

The central objection to both models is that they don't take normative powers like consent seriously enough. On neither version is a person's past authorization given any fundamental moral importance. If we know that, by interfering with Ulysses now, we can help Ulysses get what he really wants, or what he will stably want in the future, then we have all the justification we need. As both Savulescu/Dickenson and Davis admit, the moral significance of the Ulysses contract is purely evidentiary, simply conveying what a person's true preferences likely are or what his future preferences are likely to be.⁴² This seems wrong. Intuitively, an important part of what makes it permissible for me to help Ulysses remain true to his values is that he *authorized* me to do it.⁴³ The appeal to diachronic autonomy or authenticity alone cannot explain that.

4. The Authentic Authorization Model

I am now in position to set out the view I favor. The key claim is that I have the discretionary power to authorize others to bind me against backsliding on my values, but that I can't intentionally authorize someone to bind me against acting on contrary values if I've really changed my mind about what my values are or how best to promote them. Because the earlier self is invested with the power to authorize another to bind his future self only insofar as the earlier self is better in touch with his more authentic commitments, we can call this the *Authentic Authorization Model*.

⁴² Savulescu and Dickenson, "The Time Frame of Preferences," 233; Davis, "Enforcing a Ulysses Contract," 101.

⁴³ Buchanan and Brock, *Deciding for Others*, 115-117.

Like the Competence Model, Authentic Authorization privileges a person's prior aims only when there is some defect in the agent's subsequent will. But the relevant defect need not be so severe as to undermine competence; it may simply involve everyday inattention, irrationality, or weakness of will. In another way, the account resembles the Authenticity Model, since it conceives of self-binding as justified insofar as it helps a person remain true to values that remain "his own" in some sense. However, unlike the Authenticity Model, it construes Ulysses contracts as the exercise of genuine normative powers. In that respect, the position is closer to the Sovereignty Model—it's just that, on the Authentic Authorization Model, the power to bind oneself is not so unlimited.⁴⁴

To put it more carefully, the view endorses the following two principles:

Authority Principle: At t_1 , A has the discretionary power to authorize B to take legitimate and reasonably proportionate measures with the intention of binding A at t_2 against backsliding on, or straying from, a certain configuration of values, if and only if we can continue to appropriately attribute those values to A at t_2 , as well as the endorsement of the Ulysses contract itself.

Attribution Principle: We can continue to attribute a certain configuration of values that A held at t_1 to A at t_2 if and only if A's mind has not stably changed about what those values are in a way for which A is answerable.

The remainder of this essay unpacks these two principles. In this section, I make the case for thinking that we have a limited power to bind ourselves against backsliding but not against a change of mind. In Section 5, I will take a closer look at the distinction between backsliding on your values and changing your mind about them. In Section 6, I turn to the perplexing case of

⁴⁴ Authenticity is also central to the accounts in Willigenburg and Delaere, "Protecting Autonomy as Authenticity" and Michael Quante, "Precedent Autonomy and Personal Identity," *Kennedy Institute Ethics Journal* 9 (1999), 365-381. However, these authors do not explain how authenticity is related to the principal's discretionary power to authorize the deputy.

those who, like some dementia patients, have permanently lost some of their old values due to psychological impairments.

A. Sovereignty, Authenticity, and the Limited Power of Self-Binding

Let us begin with the idea that each person, insofar as she is competent, properly enjoys sovereignty or authority over her own life. This means we have strong, usually conclusive, reason to defer to a competent agent's self-regarding decisions, not because they are objectively right or subjectively rational, but just because they are *his* decisions. Recognizing this content-independent authority is bound up with the broadly Kantian notion that we should interact with others as reasoning agents with whom we stand in a relationship of basic practical equality.⁴⁵

To endorse the underivative importance of sovereignty means we reject the Diachronic Autonomy Model. Interfering with a person's present preferences for the sake of his future preferences is as much a failure to respect autonomy as interfering simply to make him better off.⁴⁶ But we also reject the pure Authenticity Model. As Garrett Cullity observes, "It can be objectionably paternalistic not just for us to force a person to comply with *our* judgments of what is good for her, but also to force her to comply with *her own* judgments of what is good for her. We need an account of the right to self-determination on which that right protects us from akrasia-correcting busybodies too."⁴⁷ The effect of sovereignty, therefore, is to screen out certain

⁴⁵ Garrett Cullity, *Concern, Respect, and Cooperation* (Oxford: Oxford University Press, 2018), ch. 2. See also Stephen Darwall, *The Second-Person Stance* (Cambridge, MA: Harvard University Press, 2006) and Arthur Ripstein, *Freedom and Force: Kant's Legal and Political Philosophy* (Cambridge, MA: Harvard University Press, 2009).

⁴⁶ Savulescu and Dickenson, "The Time-Frame of Preferences," 230.

⁴⁷ Cullity, *Concern, Respect, and Cooperation*, 188-189.

substantive reasons—whether these concern welfare, diachronic autonomy, or authenticity—from the appropriate consideration of the bystander.⁴⁸

The challenge is to hold onto this notion of personal sovereignty without getting pulled back into the orbit of the Competence Model or the Sovereignty Model—for it may seem that we either need to say (i) that we should always respect the present will of a competent agent, or (ii) that a competent agent can always exercise his present authority to forfeit his future authority.

To meet this challenge, we need to reflect further on the justification of the relevant normative powers. A person's authority over her own life encompasses both the exercise of her natural liberty (i.e., her fair portion of freedom of action) and the use of whatever powers she has at her personal disposal. But it does not follow (*pace* the Sovereignty Model) that the content or scope of whatever powers I have is set by the nature of sovereignty itself. True, the possession of some normative powers—particularly contemporaneous consent—is necessary to facilitate effective liberty of action. For example, one can hardly be free to decide with whom to have sexual relations if one cannot give and withhold sexual consent. But many powers that involve placing ourselves under obligations or forfeiting future rights are not like that. As I have suggested, just because you possess sexual autonomy, it does not seem to follow that you ought to have the power to give advance consent to sex, forfeiting your future ability to rescind it. For powers like these, we need a special argument why it would be good or important to have them.⁴⁹ Take the power to promise. Some have argued that it is valuable to have this power

⁴⁸ A similar point is developed in Daniel Groll, "Paternalism, Respect, and the Will," *Ethics* 122 (2012), 692-720.

⁴⁹ Suppose Raskolnikov, alone in his garret one night, decides to renounce forever his power to make promises, just to prove his absolute freedom. The next day he is distraught that he has done something rash. I think he can be assured he has done no such thing: he never had the authority to renounce that power, since possessing it would serve no valuable purpose.

because it enables us to foster trust or forge special bonds with others; but since it is not valuable to build relationships on wrongful acts, we cannot promise to commit wrongdoing.⁵⁰ Whatever the right story to tell about promising, my point is just that this is the kind of account we need for Ulysses contracts.

Consider things, first, from the agent's own perspective. I regard myself as possessing a set of values. "Values," as used here, is something of an umbrella term covering convictions, projects, commitments, ideals, what is deemed important or worthwhile, and so on. But, in general, to value something is to be disposed to hold certain beliefs about its worth, to treat certain considerations as reasons for action, and to be susceptible to a range of appropriate emotions focusing on that thing.⁵¹ Our values, moreover, are not just an assortment of concerns. They exhibit a degree of structure and system. Some derive from others. Some are more deeply and fervently held. Some values constrain the other values we can faithfully pursue. And some values better cohere with one another than others. Let us say that a person's *evaluative perspective* is comprised of the total configuration of his values, particularly as defined by the most fundamental and deeply held.⁵²

Since my values represent my perspective on what I have reason to do, I naturally regard backsliding on them, or non-culpably straying from them when incompetent, as bad.⁵³ But, recognizing that I am susceptible to backsliding even while I remain competent, it is reasonable to want to have techniques of precommitment that can prevent it. And, since having someone

⁵⁰ Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1986), 173.

⁵¹ Samuel Scheffler, "Valuing," in his *Equality and Tradition* (New York: Oxford University Press, 2010).

⁵² Valerie Tiberius, *Well-Being as Value Fulfillment* (Oxford: Oxford University Press, 2018).

⁵³ Henceforth, I stretch the ordinary meaning of "backsliding" to encompass non-culpably straying from one's values.

else help bind me to my values will sometimes be more effective than unilateral self-binding, it is also reasonable to want the ability to engage in cooperative precommitment. But given that I prize my authority over my life, I will want this interest protected by a power that I can exercise with discretion. Wanting the ability to bind myself to my values isn't the same as approving of others who would take this upon themselves without my consent.

But why should the bystander think it is a good thing for Ulysses to be able to bind himself to his values, given that the bystander doesn't necessarily endorse the same priorities? The answer, of course, is that the bystander recognizes the importance of "authenticity," of Ulysses living in accordance with his own settled values and aims, rather than following just any passing desire. This is not the same as saying (as Davis does) that it is a good thing for a person's *circumstances* to conform to his preferences. The present claim is that it is good for a person's *actions* to be governed by his own "governing part," his own present evaluative perspective. That is the rationale for self-binding.

This point must be handled with care. Because the bystander recognizes that Ulysses has authority over his own life, the bystander does not take himself to have *carte blanche* to bind Ulysses to his values. However, the bystander can see the value in Ulysses having a power that he can use at his discretion to give others permission to help him avoid backsliding. The function of the principal's power, then, is to partially remove the screen that prevents the bystander from concerning himself with the principal's interest in abiding by his own values. In being invited to intervene, the bystander is enabled to help the principal avoid akrasia without being a busybody.

Now consider this: Would it be valuable to possess a more unlimited power to bind oneself to a course of action, even after one has genuinely changed one's mind about what's valuable or

how best to promote one's values? Well, let's return to the agent's point of view. At first, it might seem that I would welcome having that power, since I may care more about my current values than my future ability to live in accordance with whatever I will care about then. Some philosophers think it is irrational to do things you'll later regret. But I might not care now if I will later regret what I've done. The problem, however, is that I also have to consider that it would not only be me-now that would have this power; my past self would also have had it. And it is hard to see how, from my present perspective, I could think it would be a good thing if my past self would have been able to bind me to values that I no longer endorse. So, while I could wish that my present self had an unlimited power of self-binding, I could not wish that I possessed that power at every point in time. Therefore, the unlimited power fails a test of temporal generalization.

The power of unlimited self-binding looks no better from the bystander's point of view. The bystander, as we've observed, may not endorse the principal's values. His distinctive reason for helping the principal involves the moral importance of authenticity. But if authenticity is important, then it seems like we have a powerful reason for not binding a person to former values that are no longer really his. This consideration has weight even if the later self is incompetent to make his own decisions, so long as he retains an authentic evaluative perspective.⁵⁴ So, there's really nothing to be said for deferring to the resolution of the earlier self, if the values and aims supporting his resolution are no longer attributable to the later self. Instead, the bystander should see the agent as having an interest in living in accordance with his developing views about what is valuable—what we might call an interest in “unfolding authenticity.”

⁵⁴ I expand on this point in Section 6.

B. Further Limits on Ulysses Contracts

Some further limits on Ulysses contracts should be mentioned. Ordinarily, the amount of force that may be used to bind Ulysses is set by what he (perhaps implicitly) intended to authorize. If I casually say, “Don’t let me drink too much tonight,” it’s probably understood I wasn’t giving you permission to lock me in a closet until morning. But, whatever I intended, there are also limits on what I *can* authorize. Just as I do not have an unlimited power to bind myself, I do not have boundless authority to license just any restraint or sanction (no matter how severe) to accomplish just any aim (no matter how trivial). In particular, the constraint cannot be gratuitous, and it should be reasonably proportionate to the good to be achieved. Otherwise, the power of self-binding would not be a good one to have. In making the proportionality calculation, we have to consider the inherent value of the aim, the centrality of the aim to the principal’s value system, the badness of the constraint, and whether the principal recognizes the legitimacy of the constraint either at the time of intervention or soon after. As long as the principal remains competent and capable of changing his mind, he probably will not try to impose disproportionate constraints on himself. But if the principal loses the ability to reevaluate the arrangement, the deputy will need to play a larger role in balancing the goods and evils. When severe constraints do seem justified, it may be best that they only be administered by specially trained personnel and regulated institutions to prevent error and abuse.

There are also some means of restraint that the principal cannot authorize. For example, according to existing law and commonsense morality, there are strict limits on your ability to consent to someone inflicting severe physical harm on you or ending your life. If you generally

cannot legitimately consent to such things, you probably cannot legitimately authorize someone to impose them as sanctions in case you backslide on your resolution.

Interestingly, there also seem to be some ends to which you can consent, but to which you cannot legitimately bind yourself. As I have said, you cannot bind yourself to sexual relations if that is against your will at the time.⁵⁵ It makes no difference that your future self might be suffering from lack of nerve or be incompetent to *give* sexual consent. End-of-life decisions might have a similar character. You can refuse life-saving treatment, and sign an advance directive to that effect (in case you later lack the competence to make that refusal), but arguably you cannot bind yourself to the refusal, even if your later self is akratic or incompetent.⁵⁶ I take no stand on whether that is the correct view, but I mention it to warn against assuming that end-of-life decisions are good test cases for the permissibility of Ulysses contracts in general.

Finally, note that, since there is intrinsic value in authenticity, the deputy can be justified in holding the principal to priorities that the deputy does not share. But authenticity is not of infinite importance. Once the deputy has been invited to intervene, even though authorized, he must judge for himself whether the Ulysses contract would do the principal significant harm. If Dennis asks you to bind him to a dangerously unhealthy diet, you probably shouldn't do it, even if that is what he really values. The disvalue of this aim presumably outweighs any benefits deriving from authenticity. But, again, it matters how central the aim is to the individual's system of values: it makes sense to give more weight to a person's religious commitment to asceticism than

⁵⁵ I assume sexual bondage is only permissible on the understanding that, if "no" doesn't mean *no*, some other word or sign does. For an analysis of Ulysses contracts along these lines, see Tom Walker, "Ulysses Contracts in Medicine," *Law and Philosophy* 31 (2012), 77-98.

⁵⁶ Thus, the current Canadian law on euthanasia permits advance authorization, but the patient (competent or not) can always invalidate it. See *An Act to amend the Criminal Code (medical assistance in dying)*, 2021, S.C. 2021, c. 2.

to a less central, if still intense, desire to be thin. In any case, one reason you might think the aim is bad is because you are confident that Dennis may later, after a change of heart, deeply regret that you formerly held him to his resolution. Although the fact that a person may thank us later is not a sufficient reason to hold him to his values, the fact that he *won't* later be glad you bound him affects the proportionality calculation. In this way, we can do some justice to the intuition underlying the Diachronic Autonomy Model.

5. A Change of Mind

According to the Authentic Authorization Model, you may enter a Ulysses contract to bind yourself against backsliding on your values, but not against changing your mind about what your values are or how best to promote them. How exactly should we understand this distinction? Well, if anything counts as backsliding, it is clear-eyed akrasia: recognizing in the moment what you have most reason to do, but still acting on a motivationally powerful desire to the contrary.⁵⁷ Here, it's obvious you haven't changed your mind about your values since you continue to acknowledge them. But, beyond this, things get more complicated. Take the person who, in the grip of a strong feeling or appetite, acts contrary to the values he ordinarily endorses, but which in the moment he doesn't acknowledge. Has he briefly changed his mind, or should we also consider this a kind of backsliding?

A. Rational Reconsideration

You might think that the essence of really changing your mind involves the deliberate reconsideration of the reasons underlying your former intention to bind yourself. Thus, focusing

⁵⁷ Assuming this is a real phenomenon. For discussion, see Alfred Mele, *Backsliding: Understanding Weakness of Will* (Oxford: Oxford University Press, 2012).

on psychiatric cases, Ryan Spellecy proposes that “If a patient wished to break a Ulysses contract, her health care provider should interview her and ascertain whether she rationally has reconsidered her Ulysses contract.” If she has not, the Ulysses contract remains in force.⁵⁸ This approach would help us explain why the cases of passion just mentioned should count as backsliding, not changing one’s mind.

If the point, here, is just that getting the person to reflect on her reasons will tend to reveal what she really values, I have no objection to this. But I don’t think that rational reconsideration is the actual criterion for a change of mind. First, engaging in rational reconsideration is *not necessary* for a person’s mind to change in the relevant way. Sometimes people unreflectively drop their resolutions, or their values gradually drift and evolve. The older Russian Nobleman, for example, may have never reflected on whether his conservative views are more warranted than the socialist opinions of his youth. But since the former have plainly replaced the latter, it seems to be the conservative outlook that represents his genuine perspective. This may be irrational. It may be too passive to even describe as “*changing one’s mind*.” But the Nobleman’s mind seems to have really changed nonetheless, and this seems to be what matters.

Rational reconsideration is *not sufficient* for a real change of mind either. Temptation invites rationalization, as it’s easier to act on an errant desire when it presents itself as reason. “This is really an exception to my rule,” you think; or “Maybe I was a bit hasty in making that resolution.” In retrospect, though, you see these excuses as the mere pretexts they actually were. You may

⁵⁸ “Spellecy, “Reviving Ulysses Contracts,” *Kennedy Institute of Ethics Journal* 13 (2004), 372-392, 384. I interpret “rational reconsideration” to mean that a person has engaged in a process of reconsidering her reasons. It does not require that her conclusion is actually rational—even subjectively rational. After all, since people do not always hold values for good reasons, it would be odd to say they can only abandon values for good reasons.

have *thought* you were changing your mind, but now you know that you were only fooling yourself. Surely, it's as reasonable to bind oneself against this form of self-deceptive backsliding as any other.⁵⁹

B. Stable Evaluative Dispositions

What matters for attributing values, as I see it, is simply whether a person's mind has actually changed, not whether the change was deliberate or rational.⁶⁰ Values, as we have seen, have a temporally extended and dispositional character. Thus, it makes much more sense to say, "I really appreciated how much I value being a parent last Saturday," than "I really valued being a parent last Saturday." To speak of a person's "settled values" is almost pleonastic. The real question, then, is whether one configuration of evaluative dispositions has stably replaced the other.

Our minds can change at different levels, of course. We might entirely abandon some values or adopt new ones. We might simply change the weights we assign to existing values. Or, finally, we might just adopt a new view about the best way to promote our more basic commitments. For instance, we might decide that a Ulysses contract just isn't working for us—we don't value *it* anymore. In many routine cases of apparent preference reversal, the person's stable evaluative perspective has not changed in any of these ways. Sometimes strong passions and appetites make us temporarily forget about our values or diminish our ability to be motivated by them.⁶¹ But this does not mean that the values disappear and then return. They simply become inactive or submerged for a time. The same is true in most cases of temporary incompetence. A person

⁵⁹ Thus, I also reject Radden's suggestion that what matters is whether the subject *feels* like she has changed her mind ("Second Thoughts," 794, 799-800).

⁶⁰ I will qualify this with reference to the permanently incompetent in Section 6.

⁶¹ See Aristotle, *Nicomachean Ethics*, VII.3.

acting unusually while in a state of delirium should not be understood as having changed his mind (however irrationally)—his values remain in place; they just are not presently engaged. Naturally, reconsideration of one's reasons can lead to a new stable evaluative perspective. But it's the change in a person's evaluative perspective itself that fundamentally matters, not the causal process which brings it about.

This approach delivers the right answer in cases of unreflective value change. Even if the Russian Nobleman never set out to change his mind, and never really reconsidered his old opinions, we cannot attribute his old views to him any longer if his mind nonetheless changed over the years. True, his new evaluative perspective leads him to treat different considerations as reasons, but the change from one perspective to another need not have been made for reasons.

The same considerations imply that what begins as akratic rationalization can eventually establish itself as a person's authentic perspective. Suppose Peter resolves to give away ten percent of his income to charity each month. However, before long, he begins finding excuses to justify why he can't do so *this month*. As the rationalizations become more frequent, and the moments of recommitment to his plan less so, there comes a point when we should say the resolution is not his authentic perspective any longer.⁶²

Of course, often a person has conflicting, alternating views about what he has most reason to do. For example, Peter usually takes himself to have most reason to donate to the charity. But, when it comes time to write the check, he often sincerely believes he has been too rigid about his donation schedule and that it really makes more sense to skip this month. Each of these

⁶² I owe the example to an anonymous reviewer.

perspectives is intermittently recurring, neither stably replacing the other. Can we say that either of these dispositions represents Peter's more authentic standpoint?

There is no single mark that distinguishes more authentic attitudes from less authentic ones, but there is a range of factors that, taken together, bolster the authority of a particular perspective to speak for the agent. First, judgments that we make when we are more aware and lucid have a better claim to represent our authoritative perspective, since in those moments it seems we are, comparatively speaking, "all there." Second, even though two perspectives may be of equal duration as dispositions, one may be expressed much more frequently and for greater stretches of time than the other. Indeed, if a person retains a disposition to value something, but the disposition is seldom activated anymore, nor is expected to be in the future, then this is no longer a part of the person's evaluative perspective at all. Third, values seem more representative of our authentic perspective when they are more robust in the sense of being expressed in a wider variety of circumstances and moods, as opposed to being activated only under very particular conditions. Fourth, value judgments seem more authentic when they are better integrated with our broader web of commitments and are more permeated with our deeper and more persistent emotional responses.

When these various "credentials" run in harness, it is straightforward to distinguish the values that we should treat as the agent's more authentic commitments. Even when they don't, if one perspective has considerably better credentials than the other, we can still identify it as the agent's more authoritative standpoint.⁶³ If neither has markedly better credentials than the

⁶³ Quante ("Precedent Autonomy and Personal Identity") takes a similar view, except he places more emphasis on rationality.

other, we should conclude that the person is simply of two minds. It is no part of my argument that every conflict within the self is a clash between a true self and certain besieging motives. When a person *is* of two minds, the Authentic Authorization Model says that a bystander has no good reason to side with one perspective against the other, and a Ulysses contract to bind the person to one of the two would lack authority.⁶⁴⁶⁵

C. Problems of Forecasting and Identification

If we say our minds genuinely change when one evaluative perspective stably replaces another, we seem to face a forecasting problem. In retrospect, it may be clear whether a new configuration of values has supplanted its predecessor, but how can the deputy know that in the moment? Well, he may have to act under uncertainty. But, of course, that is not unique to this context. What's important is that we can reasonably surmise whether a shift in judgment has really taken firm root or is just a fleeting attitude. This is a large subject—much of it not particularly philosophical—but a few remarks may be in order to indicate the general drift of my thinking. For example, we should ask whether the person has reversed herself in this way before. That might diminish our confidence that this will be a stable change. We should also consider how lucid and aware the person is in forming her new judgment, on the presumption that such factors make judgment shifts more stable. We can also look at the structural relations between

⁶⁴ But, if I am torn between A and B, yet value doing one or the other, then I can probably bind myself to either, just plumping for one.

⁶⁵ Someone could think that authenticity is more value-laden than this. It is apparently common to think of the “true self” as fundamentally good (see George E. Newman, Paul Bloom, and Joshua Knobe, “Value Judgments and the True Self,” *Personality and Social Psychology Bulletin* 40 [2014], 203-216). Therefore, the young Russian Nobleman may think that he is justified in binding his future corrupted self to his current values, since he regards his current values as a better representation of his true self. For my part, I think it is better to identify a person's authentic evaluative perspective in more psychological terms, without relying on substantive judgments about which perspective is superior. My thanks to an anonymous reviewer for raising this issue.

the new judgment and the subject's other values and emotional dispositions, since greater rational and affective coherence would seem to be conducive to greater stability.

Further, if it's easier to detect stability after some passage of time, that speaks in favor of letting some time pass. Ulysses contracts may, at the principal's discretion, build in a waiting period to see if an apparent change of view sticks. Some temptations and mental disturbances have characteristic time profiles, making it possible to wait them out. This is permissible because the Authority Principle restricts what Ulysses may *intentionally* bind himself against. If a waiting period foreseeably, but unintentionally, restricts Ulysses once he has actually changed his mind, that is not necessarily an objection. There is, naturally, a trade-off, here. The more we wait, the more confident we can be that a person has really changed his mind. But, if the person *has* changed his mind, the longer delay also constitutes a greater infringement of liberty. We can treat this, however, as comparable to questions about the severity of the constraint. The more important the aim, the more reasonable it is to impose a waiting period and risk temporarily hampering the freedom of a future changed self.

In some situations, moreover, we know that backsliding is more likely than a genuine change of mind, yet it remains difficult for the deputy to discriminate between the two. In such cases, the principal may want to enter, as a second-best, a Ulysses contract he cannot get out of under any circumstances. If there were an omniscient deputy, these arrangements would not be permitted. But presuming the duration of the Ulysses contract is not too long, and the costs of wrongly assuming there's been no change of mind are not too great, then a no-release contract seems proportionate and justifiable. This is like a regulation which, with a blanket prohibition (e.g., no swimming at a beach with a dangerous tide), intends to protect people against risks they

wouldn't want to take, but which, as a foreseeable but unintended side effect, prevents a few from doing what they really want. The important point is that, although the intuitive permissibility of some no-release Ulysses contracts might seem to support the Sovereignty Model, there is a way of justifying them in real-world circumstances on the present account.

But there is a deeper forecasting problem: What if it's indeterminate whether the new perspective will replace the old one, and it all depends on whether the deputy holds the principal to the Ulysses contract? Maybe Rahul goes to law school to advance social justice and tells you, his classmate, not to let him be seduced into joining "one of those soulless corporate firms." But now, the lucrative offers from those very firms look mighty attractive. Suppose you have reason to believe that, if you don't intervene, Rahul will take the job at the big firm. And, his decision being constantly reinforced by the values of his new environment, he will come to look back with satisfaction that he so sensibly changed his mind. But, if you hold him to his resolution, he will have a rewarding career representing the disadvantaged, forever grateful you helped him resist that moment of temptation. What should you do? The case for holding Rahul to his resolution is weaker than in standard cases, since the future Rahul will be living a life in accordance with his values no matter what you do. Nevertheless, if the restraint isn't excessive, Rahul can still authorize you to bind him to his current values, since (thanks to your intervention) he will never actually change his mind.

6. Answerability for Value Change

I have said we can attribute values to a person if she has not stably discarded them. But what about those who, due to an impairment like dementia, permanently lose a grip on the values they formerly held? Consider this case related by Agnieszka Jaworska:

Mr. O'Connor was a deeply religious man for whom thoughts ... of withholding lifesaving measures for whatever reason were completely unacceptable. In his seventies he developed Alzheimer's disease.... With the loss of capacity for complex reasoning, most of his religious beliefs gradually faded away. Then came a terrible emotional blow: the death of his wife. He has now begun saying that he does not want to go on, that he does not want to live.⁶⁶

Suppose that Mr. O'Connor, when competent, had left instructions that, if unable to make his own decisions, he was to receive all available life-sustaining care. At this point, let's assume, it's clear that Mr. O'Connor lacks the capacity to make medical decisions on his own. But the question remains: Should his proxies be guided by his former instructions, his current preferences, or somehow try to take both into account? Let me contrast my position with two others.

We might say that, when certain values permanently cease to guide a person's thought and action, no matter the explanation, the values simply cease to be his, and we should now disregard any instructions that were rooted in those values.⁶⁷ This *Psychological Approach* is a coherent view, but it has the surprising implication that, while licensing many Ulysses contracts, it calls into question the apparently less controversial practice of honoring traditional advance directives, which apply to patients who no longer have any clear preferences at all.

A second view says that we ought to draw a distinction between (A) cases in which the subject goes from one set of values to another (whatever the explanation), and (B) cases in which the subject goes from having a set of values to having no values at all. Those who remain valuers retain a certain standing to have their current evaluative perspectives respected as authoritative;

⁶⁶ Agnieszka Jaworska, "Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value," *Philosophy and Public Affairs* 28 (1999), 105-138, 107.

⁶⁷ See Savulescu and Dickenson, "The Time Frame of Preferences."

therefore, they cannot be bound by their former values. But when a subject no longer has values at all, he no longer presently has an authentic evaluative perspective to respect, and we should recognize instead the authority of his former wishes and values. Call this the *Status Approach*.⁶⁸ Much turns, of course, on how we conceive of valuing. If we take a demanding view and identify a person's values with his conception of what will make his life good or successful as a whole, then a mid-stage dementia patient who has lost track of the shape of his life might not count as a valuer anymore.⁶⁹ But we might also take a less demanding view. We might say that a person counts as a valuer as long as he retains the capacity to form judgments about what now seems good for him (even if he no longer has a grip on his life as a whole).⁷⁰ Or perhaps it is enough that he continues to care about things in the sense of remaining emotionally invested in them in a sustained and coherent way.⁷¹

Whatever the right interpretation of valuing, my issue with this approach lies with the idea that a person's former wishes and values can only matter once a person is no longer capable of valuing anything at all. The oddity of this view comes out most clearly if we imagine a case in which there is no conflict between a person's past wishes and current concerns. For example,

⁶⁸ It is dubbed the "threshold of authority approach" in Agnieszka Jaworska, "Advance Directives and Substitute Decision-Making," *Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), <<https://plato.stanford.edu/archives/sum2017/entries/advance-directives/>>.

⁶⁹ For this view, see Ronald Dworkin, *Life's Dominion: An Argument About Abortion, Euthanasia, and Individual Freedom* (New York: Alfred A. Knopf, 1993), ch. 8. Dworkin actually seems skeptical of Ulysses Contracts for competent subjects. He says that, since the right to autonomy protects the *ability* to live in conformity with one's values, we should always defer to the competent individual's contemporaneous decision to revoke his former instructions, even if the new decision exhibits weakness of will. See his example of the Jehovah's Witness in *Life's Dominion*, 223.

⁷⁰ Agnieszka Jaworska, "Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value," *Philosophy and Public Affairs* 28 (1999), 105-138.

⁷¹ Agnieszka Jaworska, "Caring and Full Moral Standing," *Ethics* 117 (2007), 460-497.

suppose it was once important to Mr. Hernandez to tithe to his church for the rest of his life and he has asked his children to honor that commitment if he cannot handle his own finances. Now his impairment has rendered him unable to remember or understand that resolution at all but continuing to make these contributions would not interfere with anything that presently matters to him. According to the Status Approach, whether we should honor that former commitment depends on whether he currently counts as valuing anything. But why should our view about the moral relevance of this former commitment depend on whether he continues to care about other, unrelated things? It seems more plausible to say that the values we attribute to a person should have something to do with his relation to that value, not his relation to other values.⁷²

This brings me to the view I endorse. Ordinarily, to attribute a value to a person, he must not only actually value it at some level, but he must also have some basic understanding of what he values or cares about. How demanding that requirement is depends on the nature of the object. For example, a basic understanding of a cherished relationship or a valued place is often not that difficult to achieve. A person might even have the requisite understanding without the ability to articulate it. But attributing other values to a person requires considerably more cognitive capacity. Suppose a child is swept away by the pageantry of a fascist party and thinks of himself as a follower, but really can't appreciate the significance of the ideology. In that case, I presume we wouldn't attribute to him the values of the party, even if he thinks he endorses them.

Similarly, when a person's values change, I want to say that a person must remain *capable* of understanding and appreciating the former values as they are lost or discarded, if we are going to attribute the change of view to him. Let's fill this out a bit. The person needs to be able to be

⁷² I turn to the harder cases where there is a conflict between the past wishes and current concerns below.

able to understand the content of those values on a conceptual level. Further, he needs to retain an understanding of how those values might relate to the circumstances of his life (e.g., a person should be able to understand, not just what it means to be a parent, but that he is a parent). Finally, he needs some appreciation of what he formerly thought good or appealing about those values and at least a rough sense of how important they were to him (even if he is no longer sympathetic to them).⁷³ If he lacks the capacity to understand and appreciate his former values in these ways, then the fact that he has ceased caring about what he once thought important is not something up to him, not something he is *answerable* for. In such cases, then, it seems fitting to say that the former value is *still his* in a moral sense, since we do not attribute to him the change in view. I call this the *Answerability Approach*: We should (at least presumptively) continue to attribute a person's values to him as long as he has not stably abandoned them in a way he is answerable for.⁷⁴

On this view, traditional advance directives are valid because patients who are unconscious or unresponsive never abandoned their values in a way for which they are answerable. For the same reason we say that, since a person like Mr. O'Connor has lost the abilities that would allow him to make sense of some of his former values, he never truly abandoned them, and we should

⁷³ Could a person lose the ability to appreciate some of his former values, not due to psychiatric illness or neurological damage, but simply due a change of character? This doesn't seem very probable to me. I am not imagining that the person must feel the *pull* of his former values—he simply needs an ability to understand what his former values were and what they meant to him. Even if we thought that a person's character could so harden that he couldn't appreciate what he used to value, it seems likely that there was a period in which he stopped caring about his former values when he did retain some ability to appreciate their significance. Therefore, he would have been answerable at that time for discarding the values and we would not attribute those former values to him now, even if he couldn't presently appreciate what they meant to him. My thanks to an anonymous reviewer for raising the question.

⁷⁴ Suppose we were convinced that someone's former values were entirely vicious or worthless. Would we continue to attribute those values to him, if he was not answerable for his having discarded them? I am not so sure. But I will not explore this complication here, since ordinarily a person's former values are not thoroughly bad.

continue to attribute them to him. By contrast, consider the version of the Russian Nobleman whose values simply evolve over time and who avoids considering whether he has good reasons for rejecting his former values. Though he may not actually appreciate the significance of some of his past values any longer, he presumably remains *capable* of doing so. Therefore, he *is* answerable for not holding those former values, and they are not attributable to him any longer. Finally, suppose this Nobleman later suffered from some serious impairment and was unable to appreciate either the values from his youth or from his later years. Since he was answerable for having discarded his youthful values, but not the later ones, we would continue to attribute to him the latter but not the former.

On the Answerability Approach, the fact that we might attribute certain former wishes and values to a person does not imply that we should ignore what he presently cares about. But this means that there can be conflicts between the person's past instructions and current preferences. How should we handle these? I think we should ask ourselves whether the person has undergone any new experiences that would have changed his mind about his former intentions (if only he hadn't lost the capacity to appreciate them). That sometimes answering this question will be difficult should not overshadow the fact that in many cases the answer will be obvious. For example, suppose that Ms. Richards had intended to leave an inheritance to her grandchildren. Now, living in a residential care facility, she really doesn't remember that she has grandchildren, but she has fallen in love with another resident and frequently says that she wants him to have everything she owns. I think if we ask ourselves whether Ms. Richards would have changed her mind in this way, if she had not lost the ability to recall her commitment to her grandchildren, then (if Ms. Richards is like most people) the answer will almost certainly be *no*.

Perhaps she would have wanted to leave her new friend something, but it is hard to believe that she would have wanted to cut off her grandchildren. Indeed, families make judgments of just this sort all the time. The Mr. O'Connor case can be approached in the same way. We should ask whether we think it likely that Mr. O'Connor would have changed his mind about his religious convictions after the death of his wife if he had retained the ability to understand and appreciate them.⁷⁵

As a final remark on this issue, let me just acknowledge that the argument for binding Ulysses to the mast is probably weaker in cases involving permanent impairments than elsewhere. In part, this is because many of the benefits that Ulysses contracts ordinarily confer depend on their eventually resonating with the subject in some way—and this typically isn't in the offing in permanent-impairment cases. Moreover, to be bound by instructions one does not even remember issuing may be particularly galling. Nevertheless, if a person's values play an important role in defining how we measure the success of her life, then past values may remain significant in that sense, and abiding by the instructions that express a person's values when he was comparatively lucid may remain an important form of respect.

7. Conclusion

I have defended the view that a person can bind himself into a better governed self, but not against acting freely as a changed self. The account is based on the idea that we have a legitimate interest in living a life that conforms to our deeper values about what is worthwhile and in taking

⁷⁵ If we think it is basically a toss up as to whether the person would have changed his mind, we should consider how serious the consequences are to make a mistake one way or the other. How important to the person was the former value? How important to him is his present concern? If one or the other seems much more important, then perhaps we should err on the side of protecting that interest.

steps to avoid backsliding on them. Since we also have an interest in having a say over how others treat us, we have the authority to decide whether anyone may bind us to those values. But, by the same token, because our interest lies in living in accordance with our developing conception of what's valuable, we lack the authority to bind ourselves against a change of heart, so long as we remain answerable for the change in perspective.

Let me close by connecting my discussion with another that has received attention of late. It might seem as if asking what a person would consent to is but one step removed from securing actual consent. But, as has been pointed out, they really draw upon quite different conceptions of autonomy. Actual consent is rooted in respect for another's sovereignty. Hypothetical consent, by contrast, appeals to authenticity—to what is consistent with his deeper values.⁷⁶ What makes Ulysses contracts particularly interesting, philosophically speaking, is that a proper understanding of them requires an integrated and nuanced account of both conceptions of autonomy.

⁷⁶ Enoch, "Hypothetical Consent and the Value(s) of Autonomy." Enoch's term for authenticity is "non-alienation."