Interpreting "What One Would Have Wanted"

Stephanie Beardman — 7/17/18 — 8,450wc DRAFT

Decisions involving the care and treatment of those with advanced dementia have high stakes — they can determine the patient's quality of life, and sometimes stave off, or hasten, death itself. The stakes can make these decisions particularly hard to make, especially because it can be hard to know what the patient wants. The assumption — widely promoted in both medical and legal practice, and deeply entrenched in our society — is that the primary aim of all such decisions should be to respect the autonomy of the patient as much as possible. This means that just about anybody who has dealt extensively with health professionals when caring for one with advanced dementia has heard the phrase "what your loved one would have wanted" invoked as a way of guiding decisions about treatment and care. The idea is that in such situations one should "preserve the patient's autonomy through the mechanism of substituted judgment."[1]

The official view is this.[2] If a patient has made their wishes clear beforehand, when they were well enough to do so, then those wishes should be followed; otherwise, one should attempt to reconstruct the autonomous decision that *would* have followed from the patients values and preferences had she been able to make such a decision. And if that's not possible, then one should try to do what is in that patient's best interest.

My discussion will center on the middle stage of these guidelines: on the reliance of "what the patient would have wanted" as a way of respecting autonomy and in the absence of an advance directive.[3] The standard policy applies to surrogate decision-making for those who were once competent and then either became cognitively impaired or comatose, their decision-making capacities impaired or obliterated by

[1] Hurley et. al., in Volicer, Ladislav and Hurley, Ann, eds. (2004), pp. 163-4.
[2] B & B, see also Jaworska SEP ///
[3] I leave the other elements and assumptions behind the standard policy intact in this discussion. Thus, for the most part, questions about the following issues are screened off: the normative force of advance directives and their relation to the aim of respecting autonomy;

disease, injury, or medical and other harmful events and interventions.[4] But I will focus here on those with advanced dementia — those who were formerly competent and experience cognitive impairments that render them mentally incompetent while retaining consciousness. This is important, because, for reasons that will become clear, the problems I identify with WOULD do not arise for individuals who become permanently unconscious. Unsurprisingly, the official policy works best for the sorts of cases for which it was developed in the law — those in permanent comatose or persistent vegetative states. So what I say here will not affect the viability of that policy as it applies to those cases.[5] The larger framework in which this paper is situated concerns how we should make decisions of treatment for those who have advanced dementia and are not in a position to decide for themselves. There is obviously a very wide range of types of cases as well as issues at stake. But part of the point of the present paper is that dementia (and permanent cognitive impairment) has to be treated in a special way. Some of the reasons why this is so can be demonstrated by showing what is problematic about implementing the second step of the guidelines in such circumstances.

The aim of the present discussion is to raise some problems of interpretation for the notion of "what one would have wanted" (which I'll sometimes refer to as 'the notion' or WOULD, for short). Against prevalent assumptions in current medical and social practice, I argue that the ability to know, not to mention even the very meaning of, "what the patient would have wanted" is problematic (indeed, in certain relevant contexts, it might be unintelligible). In particular, I would like to explore three potential problems with WOULD in such contexts: (1) that the notion is incoherent, (2)

---

what is good for the patient; the role of benevolence; and the role of counterfactual conditional desires (and advance directives) in determining what's in a person's best interest.

[4] The policy does not apply to those who have never had the relevant decision-making capacities, such as children or those born with conditions that prevent them from becoming competent. Note that in such cases, respect for autonomy is usually not considered the relevant underlying principle, in any case; the standard instead is that of beneficence and best interest.

[5] To ADD later in paper: I mentioned at the outset that the conceptual problems I identify with WOULD do not arise for those who are rendered permanently unconscious. What we can see now is that, since the policy does not itself explicitly distinguish these cases, this shows that the criterion of mental competence to make decisions for oneself is too blunt for adequately sorting through cases. Moreover, it suggests the threshold of competence is too high to capture normatively salient features of the lives of those with dementia. ///

that the answer to the question is fundamentally (metaphysically) indeterminate, and (3), that, in cases where the question is intelligible and its answer reasonably determinate, considering what the patient would have wanted is itself irrelevant to the matter of respecting autonomy, because what the person would have wanted does not play an independent role — it is determined by normatively independent features of the situation. Thus, when the counterfactual is appealed to as a way of respecting or preserving patient autonomy, WOULD is conceptually, epistemically, and practically problematic.

Taken together, these claims indicate that WOULD is not a useful question for surrogate decision-making that aims to preserve the autonomy of an individual with advanced dementia. This is not to deny that the notion may have heuristic value. Often, when people are asked to consider the question WOULD, they can achieve greater mental clarity or peace of mind. But just as often, and especially in difficult cases, they may be further confused by it; it's an empirical question what these effects are.

1. The scope and usefulness of WOULD

Before turning to the main arguments, it will be helpful to set aside the uncontroversial ways that would is — for our purposes — relatively unproblematic for the mechanism of substituted judgement. In this section, let's consider what makes the notion useful in many cases for surrogate decision-making.

As indicated above, when choosing what to do regarding a formerly competent person whose decision-making capacities are compromised or absent, one can make decisions that are in their best interests or for their good, without purporting to be their agent. In that case, the standard is that of the individual's best interests, and the underlying principle is that of beneficence. Alternatively, one can act on behalf of the person, that is, as their representative or agent, their surrogate. The standard is that of substituted judgment, and the underlying principle is that of respect for autonomy. The question to consider here is this. When the aim is to respect autonomy — when making decisions *on behalf of* somebody — is asking what she or he would have wanted a useful way of trying to be their agent?

To illustrate how the notion of WOULD generally works, Ronald Dworkin recounts the story of a daughter of a widow in intensive care who "said, with great confidence, that her nearly dead mother would want everything done to keep her alive as long as possible. The daughter's reason was not a remembered conversation or even an imagined one, but a family tradition that she insisted her mother shared: a tradition of fighting every other kind of battle to the end."[6] Her mother suffered multiple heart failures in the intensive care unit, and refused to consent not to be resuscitated after the next one. Finally she needed to be put on a ventilator and her daughter insisted she continue to be resuscitated even in that condition. As the daughter pointed out, even their pet cat, when he was dying, was given blood transfusions, continuing the family tradition of "fighting to the bitter end." For an example which expresses different values, consider the case of Nancy Cruzan, who was kept alive for eight years in a persistent vegetative state while her family argued, unsuccessfully, before the Supreme Court that she had a right to refuse life-sustaining treatment.[7] Her parents and friends, Dworkin surmises, might have relied on a similar sense of her personality and values in reaching the view that she would not have wanted to be kept alive in a persistent vegetative state: "They talked about her vivacity and her sense of the importance of activity and engagement; they thought that a person like that would particularly despise living as a manicured vegetable." (Dworkin, pp. 191-2)

There are many circumstances in which invoking such a notion is useful. Under normal conditions, there isn't generally a problem in relying on WOULD when a competent person isn't present, as when a house sitter decides the owner would have wanted her to sign for and accept a delivery package that arrives while they are away. Usually, when one has to make a decision without an affected party present, and isn't able to consult with them because of time or other inaccessibility constraints, the notion itself is not conceptually problematic. Nor is it in principle unknowable or irrelevant. Babysitters, spouses, co-workers, friends, even enemies and renegades, use the notion fruitfully all the time.

---

[6] Dworkin, 191; he notes that the case is described in Thomasma and Graber, *Euthanasia*, New York: Continuum, 1990.

[7] Cruzan v. Director, Missouri Department of Health, 497 U.S. 261 (1990).

The usefulness of WOULD expands beyond these everyday cases to those involving surrogate decision-making in certain medical contexts. Take the type of case for which advance directives were first developed and adopted in the law, involving the loss of capacity for consciousness: When the question is what to do in cases in which one is irreversibly unconscious, it is useful to ask what one would have wanted in deciding, say, whether to pay the doctor, agree to a blood transfusion, withdraw ventilator support, and so forth. There's no special problem of WOULD in the case of one who is comatose. Similarly, there is no intrinsic difficulty in considering what one would have wanted after one has died.

There's one more category of conceptually unproblematic case, involving preference changes over time. These are ubiquitous and figure prominently in decision-making and planning about the future. In this type of case, it is reasonable to hold constant relevant overarching values and personality traits and consider how these might interact with other values and beliefs when circumstances change. For example, often one can be confident that, say, as a poor college student one would have been in favor of a subsidized meal plan even if one doesn't care about this at all when not college-bound. Or, one might be able to know that as a parent one would have valued living in the suburbs near good school districts, even if one does not presently value such things as a childless single person who lives for late-night parties in the city. This is because there are certain values that one can be fairly confident will remain (and would have remained) stable, such as the value of having easy access to food and concern for the well-being of one's loved ones, or for the safety for one's progeny were one to have any.

Such preference changes are context dependent in the sense that they are due to mere change in point of view or context, holding (say) degree of self-interest and other general values constant. They involve predictable shifts in how the same circumstance or body of evidence is evaluated at different times due to changes in one's attachments and other things that one cares about. Certain preferences and values would change purely as a result of a shift in point of view — and independent of the experience of what it's like to occupy that perspective — such that general values such as love or self-interest, for example, would make certain considerations more salient. These are cases of evaluative shifts — changes in one's values and

preferences — that are not in principle difficult to imagine and for which the notion of WOULD is unproblematic. Asking what one would have wanted in such circumstances makes conceptual sense, and wouldn't require or depend on epistemic access to the phenomenal qualities of inhabiting that point of view or perspective. And just as forming conditional preferences makes sense for such cases, so, third-parties can also sensibly invoke WOULD in these cases, because one can at least in theory identify what these context-driven changes might be.

In all these cases, we are operating with a certain picture of the surrogate decision-making process according to which a particular decision process need not be envisaged (though in certain instances it might be — the point is not to exclude such a method, but simply to set aside additional imaginative hurdles that might needlessly distract from the main issue). To simplify the discussion — and also to give the most plausible rendering of WOULD — we can take the question to invoke what course of action would best accord with a person's personality, general values, stated desires, and so forth, taken as a whole. This is the usual interpretation; we needn't worry about what exactly we are trying to envision — whether we are supposed to imagine what one, say, "would have thought after reading a specific novel or hearing a specific argument? Or in the absence of any discussion or argument? In a good mood? Or depressed?" (Dworkin, p. 191) That is the sort of epistemic problem we can safely exclude, since we need not imagine a decision process at all. So we need not speculate about a particular counterfactual decision process. It's enough to determine what's most in keeping with their character, preferences, and general values. Moreover, it's important to make explicit that we are talking about characteristics of the person as expressed or embodied in the past, when the individual was mentally competent (thus sweeping aside another potential source of confusion about the temporal perspective of the counterfactual conditional).

Note, however, that while a person's values figure prominently when asking about WOULD in the context of the decisionally compromised, it's not always the case that such evaluatively weighty elements are relevant to all cases in which WOULD is successfully invoked and answered. Regarding, say, the babysitter's case mentioned earlier, one who accepts a delivery package on behalf of another is not necessarily

considering their deep values and so forth, but is making general assumptions about what most people would want in such a situation. However, it's a notable feature of the type of case on which we're focusing that such similarly broad assumptions are not usually available — there is too much variation in what people care about regarding their medical treatment and care. And the stakes, moreover, are typically higher than those of receiving a package; this is in part what makes such decisions complicated, as well as emotionally difficult. Lastly, to avoid a potential ambiguity regarding the scope of the conditional (and in keeping with a common understanding of the notion), the relevant tenseless counterfactual WOULD is: What (past, competent) subject S wants [to happen to her in the case of being cognitively incapacitated or compromised],[8] where the time of S's desire would be some time prior to becoming unable to decide for herself.[9]

Now, as we know, even with respect to these various examples of conceptually unproblematic cases, determining the answer to WOULD can be full of epistemic

[8] That is, the conditional is wide-scope: SW(D—>x), not: D —> SWx.

[9] This is not the only way to interpret the counterfactual. One might want, instead, to appeal to the hypothetical desires of the current *idealized*, *rational*, agent regarding the current subject who is cognitively compromised. (Thanks to SP for pressing this.) Thus we might imagine what the one with advanced dementia would want if she could think of her current situation rationally and without cognitive incapacity, and given her cognitive incapacities. I discuss the problem with idealizations like these at the end of the paper. I'll just say here that such idealizations of the current subject might generally be even more difficult to conceptualize than what the subject would have wanted in the past, when well.

It's worth pointing out, however, that, in the everyday cases mentioned above, what we naturally do is think of what one would now want, if they knew what was going on (as when one would want a package delivery to be accepted), and it would be strange to insist on what an earlier version of the subject would have wanted me to do know if she had known earlier than this was going to happen. But the situation involving dementia does not allow this simple interpretation of WOULD, because people are usually uncertain to what extent the same person remains. Nevertheless, I grant that it is possible to go with the first, syncronic, semantic interpretation even in this case. One might do so because (i) one thinks the same self has persisted, or (ii) one has shifted the relevant agent to the present one, while recognizing that he/she is hindered by cognitive problems. That is: What would this demented being want if she could be competent, or a full agent, or.... In this latter case, the same conceptual worries arise — can one imagine a self without simply collapsing that question into the self when well? Or maybe, one is imagining the self when well, but now *fully* informed about the state of dementia (say, the individual is cured of dementia for a day while retaining all memories of what it was like, and could tell us what she wants now that she doesn't have dementia, *about* her demented self). This might be like the case of the person who merely cannot express themselves, but is otherwise competent. Then we'd want to know what the person before us *now* wants, if they could but tell us.

challenges, and one might easily get things wrong, or not know what to think.[10] Those difficulties are not our topic. The deep problem with WOULD that I want to focus on is not that it can't always be obviously or easily made precise. Neither is the problem that one isn't sufficiently imaginative or doesn't know enough about the relevant facts, such as what sort of person the patient was in relevant respects, how she would have perceived the present situation, and so forth. Nor is it just that the answer might be vague or happen to be indeterminate. In sum, the problem is not that the notion is fraught with contingent epistemic difficulties, though it is. These are, in theory, resolvable. Rather, the problem I want to focus on is that, in certain contexts, WOULD may be *in principle* unanswerable. Then it wouldn't be the right metric or guide to use when one is a surrogate decision-maker for many, perhaps most, situations involving dementia, either because the notion is unintelligible, or because the answer to WOULD is necessarily indeterminate (and thus there is, in principle, no fact of the matter).

2. Conceptually problematic cases

The conceptually unproblematic cases referred to earlier share a common feature. They're all cases in which the future prospect does not include a radically different subjectivity, as with the prospect of loss of consciousness or of simply being unavailable for the relevant decision-making. The question WOULD is supposed to ask about the values and preferences of the person when well and fully competent, regarding the prospect of their own future dementia. It invokes the subjective point of view of the past (or present), competent, person when hypothetically thinking of her future (or present) from the outside, as it were (because, from her point of view, it's

[10] There is a large literature on the epistemic difficulties of counterfactuals. See, for just one example, the trouble involved in determining the relevant counterfactuals in full information and ideal theories of the good (cf. Loeb ///, + other references ///). In practice, things are worse than even the philosophical analyzes would predict. While it's difficult to asses the accuracy of judgments of WOULD, studies examining how people implement advance directives, the situations don't inspire confidence that people are getting things right. For example, empirical work strongly suggests that people are bad at screening off their own values, even when making good faith efforts to do so in acting on behalf of another. (Indeed, medical practitioners who might have never known the patients when well are better at discerning what those patients requested in explicit written directives than are close family

removed from her in time or space). As we've seen, this is not a conceptual problem when the decision-making context does not include a radically different subjectivity, as in the case of persistent vegetative state, death,[11] or that of a competent person who is unavailable for consultation about what to do. In what follows, I'll suggest that the reason that WOULD is a legitimate question to raise in such cases is that the competent self wouldn't have to imagine what the future state would be like, in order to have an informed preference, and there's no distant self with a potentially competing set of preferences and experiences to take into account.

But invoking WOULD might become especially problematic in the face of the particular condition in which a formerly competent person remains fully conscious while changed in significant ways. In the rest of the paper, we'll consider cases in which the future prospect includes a radically (and, as will become important, saliently) different subjectivity from that of the past competent self, as typically occurs in cases of advanced dementia. We'll explore three worries about appealing to WOULD in such contexts: (1) that the notion is incoherent, (2) that the answer to the question is fundamentally (metaphysically) indeterminate, and (3) that it's irrelevant.

The following two sub-sections, which argue for the first worry, raise the question whether there might be a category or subset of counterfactual conditionals that is necessarily incoherent. Each subsection, respectively, identifies two potential sources of the incoherence of WOULD (if indeed it is incoherent), and suggests an important feature of the semantics of counterfactuals: Section 2.1 considers the possibility that (i) certain counterfactuals — namely, those referring essentially to ineliminably first-personal subjective states — may be incoherent because of the transformative nature of the experiences themselves. Section 2.2 presents at a different diagnosis of the incoherence, that (ii) counterfactuals referring essentially to ineliminably first-personal subjective states may be incoherent simply because the comparative subjective states are relevantly dissimilar in a way that precludes a single global

---

members, who think they know their loved ones better (this is maybe in part *because* they think they know their loved ones better). [cf. Howard, notes 16, 17, 18, 28, 27. ///]

[11] If the assumption that there's nothing that it's like to be in these states is incorrect, then invoking WOULD would become problematic, for the reasons I go on to explain. (Though even then, in the case of death, one is usually concerned to preserve the autonomy of the

description of what is wanted. As the examples will suggest, the point here is not restricted to dementia — it's generalizable to all cases where there is a relevantly significant change in subjective state (whether the change is global, as it likely is for many forms of dementia, or narrow in its scope, as it is in the case of, say, becoming sighted).

2.1 First path toward incoherence: transformative experiences

Consider the case of a congenitally blind person who will soon have an operation that gives them sight. You now want to know what color to paint their bedroom. (Let's assume that, for various reasons, it's not possible to take the route of simply waiting for them to get the operation in order to ask them what color they'd prefer.) In such a case, it seems inappropriate to ask what they would have wanted, before their present altered state, *about* their present state. Moreover, I take it it's obvious that asking WOULD in such a situation is unreasonable not just because the answer seems indeterminate (it might not be: for example, the person in question might have had an opinion of what they would want, may be obsessed with red, or whatever), but because you cannot coherently combine their counterfactual desire with the thing you want to know — what color would they most want as a person who can actually see color? That is, any counterfactual desire would not be relevantly informed by the most relevant feature of color in this case, which is the visual experience of it. It wouldn't be *about* the right thing.

In this example, the mismatch between subjective experiences is extreme: you're trying to imagine what a non-sighted person would want for their sighted self regarding an experience of which that past self knows nothing. Something analogous might be going on in the case of advanced dementia. Because the subject retains consciousness while many cognitive abilities fall away, what it is like to be in the world changes dramatically for the subject. The fact that it might also be difficult, perhaps impossible, for such a being to conceive of the differences in subjectivity, or to even conceive of what their experience is like as such, does not undercut the point

---

living person's past self regarding their legacy and what they would have wanted to happen

that there is something that it's like — and, moreover, that that experience might be not only radically different from that of their earlier self, but also impossible to adequately imagine prior to entering that state.

The vampire example is another paradigmatic case of what's known as an "epistemically transformative experience." This technical term, introduced by L.A. Paul (2014), refers to a state that it's impossible to know the phenomenology of what it's like, prior to actually experiencing it. It refers to the effects of the subjective, first-personal nature of an experience itself on one's epistemic and psychological state of mind (that is, on one's warrant, knowledge, beliefs, preferences, values, desires, hopes, fears, and so forth) that are in principle not knowable or predictable before undergoing the experience itself. As I'll use the term, a transformative *choice* will include both the choice to undergo a potentially transformative experience as well as the conditional choice about what to do given the occurrence of a transformative experience.

It is important to keep in mind that, in the technical use of the term, "transformative experience" does not refer merely to epistemic, evaluative, or preference shifts, but rather to a state that's essentially constituted by an unknown and ex-ante unknowable phenomenology. The relevant epistemic and personal changes are a direct result of the first-personal phenomenology of the new experience. What's learned or discovered is what it is like; the personal changes occur *in virtue of* the radically new phenomenology of the experience. Contrast this with the possibility that one can change one's mind about what one wants and values, and do so radically, in a way that wouldn't necessarily count as a transformative experience in the sense intended — even if it did result in what one might ordinarily call a huge transformation.

Thus, we should distinguish cases in which the epistemic or evaluative shifts are a result of the subjective experience as such and cases in which they are the result of changes in circumstance and context that would obtain even if one did not gain any new epistemic or evaluative information — simply because, for example, certain things one already knows about (say, that being bitten by a shark is painful and

---

on earth after they're gone).

terrifying) or already values (having a happy environment for one's progeny) would apply to oneself in the new circumstance, independently of the cognitive phenomenology of what it's like. These are in the class of unproblematic cases discussed in section 1.[12] Becoming a parent, for example, brings with it many changes (in sleep patterns, preferences for buying baby supplies, and so forth) that are not changes in virtue of the distinctive "what it's like" of the experience, but are simply due to a change in perspective and circumstance, and that are, as a result, possible to predict. Indeed, there are lots of experiences people consider to be transformative that are not so in virtue of the first-person salient changes in phenomenology. (Note that I'm not ruling out the possibility that parenting brings changes in virtue of the distinctive phenomenal feel of the experience as well. Moreover, one experience can result in many epistemic and personal shifts that are not a result of the subjective phenomenology, *even as* they *also* involve such shifts that *are* such a result.) So we need to keep in mind the particular, ineliminably subjective, nature of the phenomenon in question when using "transformative experience" in its technical sense.

Advanced dementia might be epistemically transformative in this sense. Something significant, what it's like, might be learned, and values and so forth might also change as a consequence of that (making the experience not only epistemically but personally transformative, thus possibly affecting preferences large and small — from how things taste, to whether one wants to continue living in such a state). There may be no way of knowing what dementia will be like, beforehand, and thus no way of coherently forming a desire on the basis of that inaccessible phenomenology. That's because what the present situation is like from the first-person perspective would be in principle, and not just as a contingent matter of fact, epistemically inaccessible to that past, competent, self.

[12] It's worth noting that Paul also excludes being-bitten-by-a-shark cases from her account of transformative experience because, as she says, in such cases "there is no need to deliberate by cognitively modeling in order to assess the subjective value of the relevant outcomes." (2014, 28) But it's important to emphasize, on my account, that such cases are excluded because they don't count as "transformative" in the technical sense to begin with. Though they overlap (e.g. one doesn't know what it's like to be bitten, etc.), the reason to avoid them has nothing to do with what's unpredictable and in principle unknowable in particular beforehand.

I do not want to argue here that advanced dementia is an epistemically transformative experience in this technical sense, though I think it is overwhelmingly plausible that it is. But I want to say that *if* it is, then that would pose a conceptual problem for invoking WOULD. In cases involving dementia, there are two subjective points of view at different times, each inaccessible to the other, and this is what produces the conceptual difficulty. On this view, the past self, of necessity, doesn't know what it's like, and doesn't have the first-personal preferences and values of the present state. She doesn't have a good grip on what it's like, or on how it will change her preferences.[13] What one would have wanted, according to this worry, has no clear sense when the question concerns the prospect of an epistemically transformative experience.

The lesson of this section is this. The problem of transformative experience is not just a problem for individuals making decisions based on their subjective values — it's a problem for third parties, the surrogate decision-makers, when their patients and loved ones have undergone a transformative experience that coincides with a loss of decision-making capacities. And this points to an important feature of the semantics of counterfactuals: Certain counterfactuals — namely, those referring essentially to ineliminably first-personal subjective states — may be incoherent because of the unknowable nature of transformative experiences.

2.2 Second path toward incoherence: two distinct subjectivities

But let's say it *can* be known what advanced dementia would be like, at least in relevant respects. There is a second route toward the claim about incoherence.

Consider again the example developed by Laurie Paul of becoming a vampire. (Imagine these vampires drink blood created in labs, so there's no moral transformation involved.) Paul presents this as a paradigmatic case of a transformative experience, but the case is interesting even if one disagrees with that designation. There's a lot we know about vampires, after all (as we know about dementia, as well) and one might think that the particular subjective feel of what it's

[13] This doesn't mean that one couldn't have had an uninformed preference; but then, that

like is irrelevant (as one might also think is the case about dementia). Even so, I want to suggest that WOULD is incoherent as applied to these cases, merely because of the radical *difference* in the two respective subjectivities.

Let's first see how this might be the case in the vampire example. Once one is a vampire, it makes sense to talk about what the vampire wants, or would want. (I'm not suggesting that the circumstance would necessarily arise where a third party would need to know the answer to WOULD for the vampire, certainly not in the way it routinely comes up regarding those who have advanced dementia.) But it makes no sense to talk about what one, *prior* to becoming a vampire, would have wanted, given the state of being a vampire, beyond making general guesses — but these wouldn't be your desires in the past, not even your conditional desires. And it makes even less sense to talk about what one would have wanted prior to that transformative event, regarding the question of what should happen as a vampire or even, whether to continue in that state after one has been, say, bitten against their will. This is because one cannot coherently apply one's former point of view to the new situation with *its* own subjective point of view.

Now, one might plausibly know whether one would have been tempted to be a vampire or been horrified at the prospect. But *once one has actually become a vampire*, everything (normatively speaking) changes. One can't, I propose, infer from the claim that one wouldn't have wanted to be x, that one should die if one becomes x. If one is then asked to suspend the will to not be in that state, and say what one would want if one nevertheless was in x, can we (or even she, in her past state) say what she would want? The two epistemic, subjective, perspectives are inconsistent; one can't occupy them both together.

For this point, I'm not relying on the more controversial claim that one can't imagine being or becoming a vampire. One might, say, know what it's like to really desire something, and correctly assume that desiring to drink blood is relevantly similar to what it's like for one now to strongly desire one's favorite dish when hungry. Similarly, the point in this section is not that one cannot know what advanced dementia would be like. The problem here is mixing the two perspectives, each with

---

wouldn't be relevant to answering WOULD.

its possibly very different global subjective experience with its attendant, and possibly incommensurable, preferences and inclinations. WOULD was a legitimate question to raise in the unproblematic cases because either there was nothing that it's like to be in those states, or, what it's like was not relevantly different from the counterfactual individual in question.

Intelligibility, of course, comes in degrees. My point here is that the coherence of WOULD is tied to how closely the self resembles itself. There are cases in which the question might be coherent, because the two selves are not that far apart. Even then, as I'll argue in Section 4, even if WOULD is at least minimally coherent, the features that make WOULD intelligible also make it the case that there's no fact of the matter about the answer to WOULD. But when a radically different consciousness emerges (whether it's transformative and ex ante unknowable or not), the worry is that WOULD is, for most cases involving advanced dementia, in principle unintelligible.

3. What about the rationality of advance directives?

At this point in the discussion, it might seem that a dilemma is looming. On the one hand, it might appear that what I've argued for so far casts doubt on the rationality of planning for contingencies that involve radically different subjectivities. On the other hand, if such planning can be rational, then why can't we simply imagine *that* decision procedure to provide an answer to WOULD? I'll respond to the first horn of the dilemma in this section. The second will be addressed in the next section.

The possible incoherence of WOULD described above might suggest the following worry about the rationality of first-person prospective decision-making about the prospect of dementia. If WOULD is unintelligible for either of the reasons proposed, wouldn't those very reasons pose an obstacle to rationally making decisions about one's future care, as people commonly do? Regarding the first path (section 2.1): If advanced dementia is a transformative experience, then how can one rationally make decisions about how one should be treated?[14] And regarding the second path (section

[14] Paul (2014) argues that transformative experiences pose a problem for standard decision theory. I do not want to argue for that position here. In this section, I present a in condensed form a sketch of my proposed solution to the problem that Paul raises.

2.2): Why doesn't the inability to merge the relevant perspectives pose a problem? Before I try to explain why the worries described above should not cast doubt on our ability rationally to plan for contingencies like having dementia, let me note here that I do not want to argue that such planning must turn out to involve conceptual confusion (though of course some of it might). I would consider it a reductio of my view if it turned out that all such advance planning is necessarily irrational.

To put the puzzle in the starkest terms: Sometimes, the question of what one would have wanted seems easily and straightforwardly answered by what one in fact *did* want. Many people talk about what they would want. A growing number write up and sign advance directives, legal documents that specify the type of medical care one wants to receive in the event one is not able to make such decisions for oneself. These include living wills, which specify the kinds of medical treatments one would like under various circumstances, as well as health care proxies, or durable powers of attorney for health care, which authorize another to make medical treatment and care decisions on one's behalf. But if advanced dementia is a transformative experience, then how can one rationally make decisions about such things? How should we take our future preferences into account in deciding what to do now, and to what extent is it rational now to make decisions that conflict with the preferences of one's future self? The prevalent view, promoted by health care and legal practitioners, is that the rational response to this question as applied to the prospect of dementia is to think about what is important to you, how you would like to be treated in various difficult scenarios, and make your desires explicit in an advance directive. The assumption is that the decision should be based on one's personal values and desires, and that the concern is what is called prudential, or egoistic (allowing, of course, that one's deepest concerns might be what one takes to be morally required). This is widely taken to be a decision that is up to the individual to answer for herself or himself.

Am I claiming that those legal documents are necessarily incoherent or unknowable? No. Of course not. In brief, my response to this horn of the apparent dilemma is this. I think that such decision-making can be rational, and yet I also think that the rational solution for making transformative choices cannot yield an answer to the question of what one would have wanted. I believe that even if WOULD is problematic for the reasons outlined in section 2, that wouldn't itself imply anything

troubling about the conceptual underpinnings of planning for the possibility of dementia. To see why, I want to briefly sketch a positive account of the rational basis of such planning. (I argue for this account in a "Prudence As a Synchronic Principle" and "The Very Present Self", in progress.) Commitment renders what one would want coherent *and* knowable. But, as I argue in section 4, we can't rely on counterfactuals about commitments to give determinate content to the question of what one would have wanted.

Let's consider Paul's proposed solution to the problem of transformative choice, and why it won't help us in the context of dementia. For any transformative experience x, Paul argues that one cannot rationally decide whether to embark upon or to avoid it on the basis of whether one wants to experience x, since one cannot know enough to assess this question. Nor can one decide on the basis of whether one wants to find out what it's like to x solely for the sake of finding out (as one could reasonably do for experiences in which not that much is at stake, like tasting something new). Rather, she argues, the answer lies in going up a level and asking whether one wants to discover the preferences one will form as a result of undergoing x, independent of whether the experience will turn out to be good or bad. One needs to choose "on the basis of preference revelation" (p. 121), that is, depending on whether one wants to discover who one will become as x unfolds.[15]

This answer, that one decide on the basis of the subjective value of discovering who one will become, will not in general work for the prospect of dementia (nor does Paul suggest it would; given a choice between developing dementia and continuing life without it, we can agree it would be rational to avoid dementia no matter what it's like[16]). It's safe to assume that no one would choose to find out what it's like to be

[15] "If you choose to have the transformative experience, to choose rationally, you must prefer to discover whether and how your preferences will change. If you choose to avoid the transformative experience, to choose rationally, you must prefer not to discover whether and how your preferences will change. [...] You choose to discover, then, the revelation involved in discovering what it is like to live life as a different kind of being, and by extension to discover whatever core preferences it is that you'll end up having." (Paul, p. 118)

[16] Cf. Paul on the rationality of avoiding swimming with sharks. pp. 27-8. But just like those cases, that doesn't settle what should happen, or what one might want, should those scary and unchosen events transpire.

irreversibly demented just to find that out. But more importantly, and more generally speaking, her proposed solution won't work for transformative cases that are not a matter of choice. It won't work for counterfactual conditionals whose antecedent is not a matter up for choice.[17] Whatever one thinks of Paul's solution for cases in which one can decide to choose whether or not a transformative experience occurs, as one ordinarily does when one decides to become a parent or to join the military, the situation changes if instead the future is not up to you, in the sense that undergoing the future experience itself is not a matter of choice. With an advance directive, of course, one is not choosing whether or not to experience dementia; rather, one is choosing what to do *if* one becomes demented. In this sense, it's similar to the prospect of facing incarceration, being an unwilling participant in a war, or getting accidentally pregnant. Like those experiences, dementia itself isn't chosen; in this, it differs from many transformative experiences.[18]

The solution I propose for making a transformative choice rationally relies on the present, first-personal, subjective perspective, yet doesn't need to know what it's like at another time (whether in the past or future). There's a way of assigning subjective value to a future transformative experience that has nothing to do with future subjective phenomenology, yet doesn't involve abdicating the first-personal perspective. One can do this by recognizing that such choices require a notion of commitment that is a product of the first-personal agential conception of deliberation.

On my view, transformative choices are best understood as essentially involving commitment. Commitment is needed when one reflectively decides to get married, become a parent, enter a religious order, or attend a conservatory. These options are pursued in the face of not knowing what things will be like, and this fact is something

[17] As Monique Wonderly reminded me in correspondence, it also may not work for cases in which there isn't sufficient psychological continuity between the individual before, during, and after the transformative experience. In the case of advanced dementia, it's unclear to what extent one could learn or discover who one will become, since the cognitive faculties necessary to make such a discovery are compromised. That is certainly possible; whether it's true of a particular case will often be an open question, since the mechanisms for communication may be independently compromised.

[18] Note, however, that the alternative picture I present below is not restricted to conditional decision-making involving cases like those in which one is deciding what to do if things turn out badly. It's intended to generalize to all transformative choice.

that the reflective agent is explicitly aware of when facing such decisions. That's the point and need for the commitment, to carry you through, to make it possible to throw yourself into the unforeseen and unforeseeable parts, and it's why these decisions are not usually taken lightly — everyone knows they are potentially 'life changing' decisions. This is true even for conditional commitments, as well as for decisions that bind your future self (such as contracts, and advance directives). If I'm captured in war, then _____. If I get pregnant, I will _____. If I get dementia, I want _____. On this model, normatively speaking, it doesn't necessarily matter what life will be like in the future, because one decides to commit presently to a course that excludes, or includes, that future whatever it might turn out to be like. This is the core of what's involved in determining who one is and will be — choosing the kind of person one wants to become or to continue being. At its core, creation, freedom, autonomy, and authenticity, these things are only possible in the face of ineliminable uncertainty. On this account, the unknowability is actually essential to the capacity to decide what kind of person you are or will become.

4. Epistemic worry: the answer to WOULD is indeterminate, and thus in principle unknowable

One might think the proposal briefly sketched presents a way of solving our initial problem of WOULD. If it's possible to make rational, authoritative choices concerning one's future demented self, then perhaps the notion of commitment at work here could serve as a sufficient basis for rendering WOULD a legitimate enterprise. Couldn't one coherently ask what one would have wanted with the picture I've presented here in mind? The answer I think is no. The proposed account cannot be used to make sense of WOULD; it won't help us to make counterfactual conjectures about what-one-would-have-wanted.

In brief, the picture I want to suggest, rooted as it is in the immediacy of agency itself, makes it impossible for others to get a grip on WOULD in the absence of one's actually having made a decision. Note that this isn't the case for ordinary (non-transformative) decision-making, of the sort that's hypothesized in the conceptually unproblematic cases of WOULD. The view of what might make transformative choice

rational is different from what makes non-transformative decisions more generally rational; one could apply WOULD for the latter sort. For example, your speculation that I would want you to shovel my driveway in the snow is nothing like speculating about what I would want were I to face a transformative experience. (There may be some analogies here to the reasons why you should not speculate what promises I would make, in the absence of my actually making them.) Knowing me as you do, you are entitled to your confidence in my standing desires; but not to knowledge of my transformative choices, choices about unknowable futures, in the absence of my actually expressing them.[19]

One might object that some commitments are predictable. I don't deny that they are. The distinction presented earlier between evaluative shifts that are predictable and those that aren't is relevant here. You can predict that I would want a limb amputated if that's required to save my life. Or that I will be committed to not swimming with sharks. But are you warranted in believing that I would, say, actually go through with marrying someone (even if you know that I and my lover want and value marriage, are in love, etc.), or that, say, I would invest all my life savings into a new business venture (even if you've seen how deeply envious I am of those who've done the same with their life savings, etc.)? Well, perhaps you can.[20] Nevertheless, hypothetical commitments, however plausible, function *normatively* like hypothetical directives: they don't have the right sort of normative force to govern surrogate decision-making. They can't make impermissible acts permissible, as actual directives can.[21] Another way to put the point — and this has everything to do with the essential

[19] It is sometimes said that not making a decision is the same as making one. But here, regarding transformative experiences, we see that not making a decision in such a scenario is *not* the same, normatively speaking, as actually making a decision.

[20] Perhaps one would be able to make such a prediction if, say, each year for a number of years, a woman has married a new person. But then it looks to me as if there is something like a habit or standing policy at work. No matter. The point I want to emphasize is not that one can never know with reasonably high certainty the commitments one might enter into, but that such hypothetical commitments don't have the right normative force to guide surrogate decision-making. Thanks to Dave Chalmers for pressing the problem and Monique Wonderly for independently suggesting a solution, thus helping me to clarify this point. The point is not an epistemic one, but a normative one.

[21] See Dworkin, R. "The Original Position," in N. Daniels, ed. *Reading Rawls* (New York: Basic Books, 1995), 18, and Howard, D. "The Medical Surrogate as Fiduciary Agent," *The Journal of Law, Medicine & Ethics*, 45 (2017): 402-420. [/// See also Enoch]

role of the unknowable in making commitments — is that commitments are often not a way of indicating what one would have wanted in specific situations... since one is committing to a course that one is acknowledging, by the very nature of the commitment, may involve elements that it's not correct to say one explicitly wants.[22]

Even being-bitten-by-a-shark cases can be overridden by an actual decision for it. This points to a further reason to think that one cannot justifiably rely upon counterfactuals about (certain) commitments. Let's say people have the right to make mistakes that harm themselves. I believe this is true in real life. (Let's assume that's right. If necessary to make the claim plausible, feel free to add 'up to a certain point'.) The principle of respect for autonomy requires that we allow this, at least to some extent. Now, is it also true for counterfactual decisions? No. Say you have reason to think that an individual would have made a decision that is contrary to their own interests (perhaps because they are 'always' doing such things). Does that give you any reason, even one that might be overridden by other considerations, to help satisfy that ill-conceived *counterfactual* desire? No. And yet, within reason, we do feel the pull to respect people's actual, informed decisions: "It's their life," we say.

This leaves us with the second potential worry about WOULD. Even in contexts in which WOULD is intelligible as a question, we can't rely on counterfactuals about commitments. In the absence of actual commitment, WOULD works as a guide only in contexts that do not involve a potentially transformative experience or a radically altered subjectivity. We shouldn't speculate about WOULD in the absence of a commitment, when in such contexts, *not* because there's a high chance of getting it wrong, but because commitments are special things that can't have normative force unless actual. The second worry implies that, in the relevant circumstances, it's necessarily futile to ask what one would have wanted and so we're not warranted speculating about the answer to what one would have wanted in such cases. So even if the notion is intelligible, we should take seriously the claim that there's no fact of the matter about what one would have wanted, absent an actual decision on their part.

[22] Indeed, it seems most people who have signed advance directives do not want them implemented literally: they want surrogates to have considerable leeway about what to do [///Howard note 25] This raises the possibility, as Dana Howard (2017) emphasizes, that people *may* often be exercising a positive will in not filling out advance directives, if they think that such documents will restrict the choices available to the surrogate.

It may be worth pointing out that the fact that it's not reasonable to speculate about the relevant types of commitment doesn't mean third-parties have no moral authority to act as surrogates. They can make decisions on the basis of the subject's interests, what's good for them, and so forth. In this case, the surrogate would take a custodial approach to the well-being of the individual, and not one of attempting to either enact or represent their will.[23] Perhaps we should take seriously the possibility that the best model is to move to the standard of beneficence rather than the standard of autonomy (especially since I think that beneficence will in many cases involve taking seriously the will of the mentally incompetent). But I cannot argue for this here, and am happy to leave the issue open in the present discussion.

5. Practical worry: WOULD is irrelevant in cases in which it is both coherent and decidable.

The third worry is that WOULD is irrelevant. In this section, I'll consider some reasons why that might be the case. In doing so, I'll address a slightly different version of the second horn of the dilemma described in Section 3 above: If advance care planning can be rational, then why can't third parties simply try to simulate such decisions on behalf of the decisionally incompetent?

In the brief sketch of prudential decision-making offered above, the element of freedom involved in agential commitments led me to emphasize the fact that, for the most part, such commitments are necessarily unknowable before the fact. On that view, we don't need reasons rationally to commit — we decide, say, on the existential model of radical choice, the sort of person we want to be (or not be), or the sort of life

[23] Howard (2017) calls this the Pure Agency Approach and the Custodial Approach to surrogate decision-making, respectively. She proposes a third approach, the Fiduciary approach, that presents a third model and focuses on the decisional capacity of the surrogate to represent the will of the individual in a way that allows for changes of mind; on her model, one can function as a surrogate agent, and use the method of substituted judgment (rather than the best-interest principle) in a way that allows for one to respect one's autonomous will while not following the advance directives. Such an approach is justified in part by the fact that most people want to allow for a lot of leeway in how their surrogates are to implement their directives, and intend primarily to render certain acts (such as the cessation of medical treatment) merely permissible, and not mandated. In that case, advance directives would not properly be understood (as I'd put it) on the model of pre-commitment.

we want to have (or not have). However, I also granted that some commitments might in fact be predictable. So I do not deny that we could commit for good reasons (and in any case the question is too large to settle in this paper). Nor have I argued that all such decisions must involve commitment in my sense. And I grant for now that even if they do, it may sometimes be reasonable to think these are predictable. But if one can have decisive reasons to commit, and if there are considerations that bear on the rationality of these commitments — or, alternatively, that make it knowable that one would likely choose irrationally — then those very considerations and deliberative rationality might be discoverable by a third party, if they know the person well enough. These would seem to provide a very good basis for answering WOULD. In that case, WOULD would be neither incoherent, nor its answer indeterminate.

But if that's the case, those reasons would be independently sufficient for guiding a surrogate — one wouldn't need the extra step of identifying what the individual would want on the basis of those reasons (and if one did need that step, then that hypothetical preference wouldn't be able to play the requisite normative role to ground respect for autonomy, as already argued for above).

Moreover, in this scenario, it's not clear in what sense one would be preserving *autonomy*. If it's obvious what one *should* want, then those considerations — that is, those reasons for what one should want — can tell us what to do independently of what one would have wanted.

Here we can expose a puzzle that's been lurking in the background. From the outset, we've been working with the most plausible — and widespread — understanding of WOULD: Rather than take it to require an earnest reconstruction of a particular hypothetical decision process, we accepted the standard reading of it, which identifies and applies relevant general values and personality traits. This has the advantage of not introducing clearly irrational reasons, such as those that would be introduced if we were relying on realistic counterfactual decision processes (since, often enough, people act for reliably bad reasons even by their own lights). Sometimes, we can know that one wouldn't have chosen in keeping with their own deep values — perhaps because of fear, neuroticism, self-sabotage, ignorance, or whatever other predictable causal influence. But the closer we get to identifying what

the patient *should* have chosen (regardless of how imperfectly they actually *would* have chosen), the less we can be said to be respecting autonomy (at least, on the common (Millian) conception of autonomy according to which it's essential that one has the freedom to choose badly). The standard policy asks us to make certain idealizations to make the counterfactual come out right.

The standard policy described in the introductory paragraphs presents the second step — appealing to WOULD — as a proxy for having an advanced directive, which would be even better as a guide toward decision-making, and thus is given lexical priority. However, the answer to WOULD can come apart from what one would have put in an advance directive. Indeed, one indication that something is wrong with giving WOULD this role is that it can come apart from advance directives. There are cases in which it's most plausible to think that, had one been able to see how things played out — say, how happy one might be as a basically healthy person with a certain type of advanced dementia — that one wouldn't have wanted the directive to be followed. This raises, to my mind, serious doubts about the lexical priority of the standard policy guidelines for surrogate decision-makers.[24]

If I'm right to raise the doubts I have about WOULD, one natural way to modify the standard policy would be to maintain the aim of respect for autonomy by following advance directives as well as possible, and rely solely on the criterion of best interests if there is no advance directive. (Note that switching to the criterion of well-being, or best interests, from that of the elusive respect for autonomy, does not itself rule out the possibility of honoring an individual's will, nor of doing so precisely because it is her will (even if that will is not authoritative, because they are medically incompetent). It is possible to act in accordance with a person's will because they will it solely because doing that *is part of* that being's own good.[25]) But a well-grounded

[24] Howard (2017) also argues against the lexical priority of the standard policy. /// though I draw different conclusions than she does///

[25] See Groll (2012) for a defense of this point. Because they lack the relevant decisional capacity, by definition those with advanced dementia do not have a present right to make decisions concerning their medical treatment. (They do have other rights.///) But that does not mean that what they will is not *in virtue of being willed* by them, part of their good. And so a surrogate decision-maker acting from beneficence will heavily weigh the person's will, and even decisively so... ///Ck. Groll 2012, FN 20, 29, 17, 24 /// There is the further phenomenon: many medically incapacitated patients think (incorrectly) their wills *are* authoritative. In such cases especially, a wrong seems done to them if one ignores their will or

answer to the question of how my conclusions in this paper should be incorporated into policy would have to await a fuller examination of all aspects of the policy.

considers it irrelevant. But what wrong is it, if they don't have a right to make their own decisions? I suggest the wrong done is that of not recognizing something of intrinsic value: acting as one wills.

Note also, however, that the question of whether and under what circumstances it's morally permissible to override an advance directive — and the sort of reasons for doing so — is a separate question.