Welfare, Autonomy, and the Autonomy Fallacy

Dale Dorsey

ABSTRACT: In this paper, I subject the claim that autonomous choice is an intrinsic welfare benefit to critical scrutiny. My argument begins by discussing perhaps the most influential argument in favor of the intrinsic value of autonomy: the argument from *deference*. In response, I hold that this argument displays what I call the 'Autonomy Fallacy': the argument from deference has no power to support the intrinsic value of autonomy in comparison to the important evaluative significance of *bare* self-direction (autonomous or not) or what I call 'self-direction *tout court*'. I defend the claim that the Autonomy Fallacy really is a fallacy, and show that my examination of the argument from deference has wider reverberations. Once we clearly distinguish between autonomy and self-direction *tout court*, it becomes much less plausible to say that autonomy of itself is an intrinsic welfare benefit.

Human beings—for the most part, anyway—are autonomous creatures. And it's a good thing, too. The exercise of autonomy is plausibly essential to much of what is valuable in life: the fulfillment of our desires, plans, and projects, and the pursuit of our interests in a prudentially rational manner. Without autonomy, we would be unable to shape our lives in a way we see fit, or—if that is different—in a way that is best for us.

In addition to the obvious *instrumental* value that autonomy provides, some have come to view the exercise of autonomy as *intrinsically* valuable, a *per se* welfare benefit in its own right. In this paper, I challenge this view. I argue that the intrinsic value of autonomy is only plausible when we confuse the value of autonomy with the value of other related, but distinct, features of a life well-lived.

The plan of this paper runs as follows. In the first section, I discuss two preliminary matters, including the precise claim about well-being I wish to dispute, and the nature of autonomy under discussion here. In the second section I discuss perhaps the most influential argument in favor of the intrinsic value of autonomy: the argument from *deference*. In response, I hold that this argument displays what I call the 'Autonomy Fallacy': the argument from deference has no power to support the intrinsic value of autonomy in comparison to the important evaluative significance of *bare* self-direction (autonomous or not) or what I call 'self-direction *tout court*'. In the remaining sections I defend the claim that the Autonomy Fallacy really is a

fallacy, and show that my examination of the argument from deference has wider reverberations. Once we clearly distinguish between autonomy and self-direction *tout court*, it becomes much less plausible to say that autonomy of itself is an intrinsic welfare benefit.

1. Preliminaries

Before I assess the comparative intrinsic value of autonomy, it is helpful to discuss a few preliminary matters, including the precise thesis I wish to discuss here, and the nature of autonomy as it is used in this discussion.

1.1. VAT

First, I want to make clear the view I take myself to be arguing against. Importantly, there are many ways autonomy and its exercise might be intrinsically valuable, only some of which I oppose here.

In particular, one might believe that autonomy, or an autonomous life, might be intrinsically valuable *on the condition* that this exercise of autonomy is *desired* or otherwise *valued* by the person in question.[1] This proposal might be a natural outcome of a desire-satisfaction or otherwise subjectivist characterization of welfare.[2] This view is not the topic of my discussion. I am perfectly happy to accept that *for some*, the exercise of autonomy might be intrinsically valuable insofar as it is important to or valued by them.

A further view, however, treats autonomy's intrinsic value as independent of anyone's pro-attitudes or states of valuing. Consider:

> *Value of Autonomy Thesis* (VAT): autonomy is intrinsically valuable for x in a way that does not depend on its being valued by x.

VAT is not exactly a chart-topping thesis concerning prudential value, but it has, nonetheless, a number of influential adherents. Steven Wall argues that

> [a]utonomy is an intrinsic value. It is intrinsically good for people to make their own choices about how to lead their lives. It is intrinsically good for them to adopt and pursue projects, not because others have tricked or coerced them into adopting or pursuing them or because they have no other worthwhile options to choose from; but because, according to their own lights, the pursuits are worth adopting and pursuing. More strongly,

> autonomy is not just one intrinsic value among many; it is one of special importance. For most people it is, or so I shall argue, a central component of a fully good life. However well their lives may go, if they do not realize this ideal to some substantial degree, they will fail to live a fully good life.[3]

Along the same lines, George Sher argues that: 'autonomous lives are, all else being equal, far better than nonautonomous ones.'[4] Will Kymlicka writes that 'my life only goes better if I'm leading it from the inside, according to my beliefs about value.'[5] Indeed, some believe that the following passage indicates just such a commitment from none other than John Stuart Mill:

> He who lets the world, or his own portion of it, choose his plan of life for him, has no need of any other faculty than the ape-like one of imitation. He who chooses his plan for himself, employs all his faculties. He must use observation to see, reasoning and judgment to foresee, activity to gather materials for decision, discrimination to decide, and when he has decided, firmness and self-control to hold to his deliberate decision. And these qualities he requires and exercises exactly in proportion as the part of his conduct which he determines according to his own judgment and feelings is a large one. It is possible that he might be guided in some good path, and kept out of harm's way, without any of these things. But what will be his comparative worth as a human being?[6]

In addition, David Brink suggests that the value of a particular choice depend not simply on its being the product of desire, but also a product of autonomous choice.[7] Though these views are substantively different in many respects, they all insist on the intrinsic value of autonomy in a way that is not at all dependent on the pro-attitudes of those whose lives are so improved.

An ambiguity in VAT should be ironed out here. Note that Sher holds that 'autonomous lives' are more valuable than non-autonomous lives. Kymlicka also focuses on the value of an autonomous *life*. Mill and Wall seem to focus on the individual autonomous *choice*. Given this ambiguity, how should we understand the *bearer* of autonomy's value? Should we value individual autonomous *choices* or only autonomous *lives*? Or some combination?

I'm going to remain neutral in this paper on the proper bearer of autonomy's value, i.e., whether an autonomous life or choice is itself intrinsically

valuable. But my discussion will focus on autonomous *choices* for the following reason. Even if autonomous lives are the proper bearers of intrinsic value, there is or must be an important relationship between autonomous choices and autonomous lives. In particular, the extent to which a life is autonomous is straightforwardly determined by the extent to which the choices in that life are autonomous. To put this slightly more technically, the autonomy of a life supervenes on the autonomy of the choices therein. And hence even if the autonomous life is the bearer of intrinsic value (*vis-á-vis* autonomy) this does not mean that the autonomous choice doesn't have an important role to play in determining the welfare value of a life, viz., *by determining the extent to which a life is autonomous*. This is not to say, of course, that all autonomous choices are equally evaluatively significant (however their value is construed; more on this in §4.4). It is merely to say that the value of autonomy supervenes on the nature and character of the autonomous choices displayed in a life. This is true whether we regard autonomous choices or autonomous lives as intrinsically valuable.

VAT is worth investigation. First, as a thesis about human well-being, it has powerful implications concerning the method by which we assess the quality of a life. But beyond its significance as a thesis about human welfare, VAT plays a significant role in moral and political argument. Kymlicka writes:

> But while we may be mistaken in our beliefs about value, it doesn't follow that someone else, who has reason to believe a mistake has been made, can come along and improve my life by leading it for me, in accordance with the correct account of value. On the contrary, no life goes better by being led from the outside according to values the person doesn't endorse... Individuals must therefore have the resources and liberties needed to live their lives in accordance with their beliefs about value, without being imprisoned or penalized for unorthodox religious or sexual practices, etc. Hence the traditional liberal concern for civil and personal liberties.[8]

Similar claims are made by many others,[9] including Mill.[10] Hence the intrinsic value of autonomy appears to be an important feature of one standard defense of the traditional liberal conception of justice, which places great importance on non-interference of the lives of citizens, and allowing a broad range of 'experiments in living', free from state molestation. If this is correct, VAT is worth our careful investigation, independently of its significance merely as a thesis about human welfare.

1.2. What is Autonomy?

The term 'autonomy' is, in the words of Nomy Arpaly, overworked. Many different concepts have been designated by this term, to confusing effect. In the current discussion, it is important to distinguish the capacity for autonomy, and the *exercise* of that capacity. Doing so is important to VAT *qua* thesis about life quality. VAT would be extraordinarily implausible if it held that *merely* the possession of a relevant capacity was intrinsically good.

But then what is the relevant capacity such that its exercise is, according to VAT, an intrinsic prudential benefit? At its heart, the relevant capacity is for a certain sort of control over one's motivations. For Arpaly, 'Agent-autonomy is a relationship between an agent and her motivational states that can be characterized by the agent's ability to decide which of them to follow: it is a type of self-control or self-government that persons usually have and that nonhuman animals do not have.'[11] Of course, merely having control over one's first-order motivations is not enough; the person who decides which desire to act on by flipping a coin has not acted autonomously.

The precise nature of the relevant form of control is a matter of substantive disagreement, even among those who accept VAT. Some insist that the right kind of control crucially involves *rational deliberation* about one's first-order motivations. (Call this the 'deliberative' model of autonomous control.) For instance, Gerald Dworkin writes that a capacity for autonomy involves 'a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values.'[12] David Brink insists that the value of acting on a desire requires that the desire is the 'product of deliberative endorsement'; this deliberative endorsement must be construed as 'an *historical* condition. Is the desire one which was produced or is sustained by a suitable kind of deliberation?'[13] Indeed, Sher notes that a plausible considered judgment holds that 'agents are not autonomous unless they have subjected their ends to rational scrutiny.'[14] Similar claims are made by, among others, Hurka.[15]

Not all partisans of VAT accept the deliberative model. Wall, for instance, does not insist that autonomous agents are required to exercise the capacity for rational deliberation in the way noted by Brink, et. al.[16] The—as I shall call it—'non-deliberative model' requires that autonomous choices are made at least in part on the basis of, or *from* one's higher-order[17] evaluative states (such as evaluative beliefs, a conception of the good, etc.) whether these first-order states are subjected to some inward focus of ratio-

nal scrutiny or not. What does it mean to act *from* a higher-order state? Importantly, this higher-order state cannot simply operate counterfactually or hypothetically. If I decide to ϕ but *simply* on the basis of a first-order desire or motivation, this will not count as an autonomous choice on the non-deliberative model *even if* it's the case that, in fact, I had the disposition to value the act in question or the goods at which the act aims. Autonomy cannot be assigned to a choice *via* this kind of counterfactual endorsement. To act from a higher-order evaluative state is for that evaluative state to at least in part explain, or take a causal role in the production of, the act in question. Indeed, this reflects the concept of autonomy generally: if I drift through life, never having chosen on the *basis* (or as a result of) of my beliefs about value, say, I have failed to live autonomously despite the fact that I may in fact possess the disposition to value the act. In this sort of case, the fact that I acted in a way I *would have* endorsed (as a result of, say, my evaluative beliefs) is a lucky accident, not evidence that I have acted autonomously. Kymlicka suggests the non-deliberative model: for Kymlicka, autonomous lives are led 'from the inside, according to my beliefs about value'. This does not require any form of rational or internal mechanism or process of deliberation but it does require, at the very least, that choices are made on the basis of higher-order evaluations. Wall suggests a similar view.[18] In addition, though he seems to find the deliberative model plausible, Sher's official account of autonomy is compatible with a looser, non-deliberative model of control: 'autonomous agents are self-directing in the more stringent sense of exercising their will *on the basis of good reasons*... we can make the most sense of what most of us believe about autonomy—our disagreement as well as our agreement—if we adopt the hypothesis that in this context 'self-directed activity' means 'activity that is motivated by an agent's appreciation of reasons provided by his situation.'[19]

The fact that partisans of VAT accept different models of the nature of autonomy suggests something of a challenge for those wishing to argue against it. One way to proceed would be to argue in favor of either the deliberative or non-deliberative model as the best account of autonomy and show that the presence of autonomy according to this model is not intrinsically valuable. This is not my strategy. I will argue that lives that are lacking in *either* the non-deliberative or deliberative model of control are no less valuable than equivalent lives that maintain *both* the deliberative and non-deliberative models of autonomous control. If this is correct, VAT fails regardless of the precise account of autonomy on offer. This argumentative strategy will be reflected in the arguments I present throughout.

In the next section, I consider an important argument for VAT, and ar-

gue that it rests on a mistake: a confusion between autonomous choice and non-autonomous, but nevertheless self-directed, choice. I then argue that this mistake is much more widespread. Without this error, VAT is much less plausible than might initially be thought.

2. The Argument from Deference

The most powerful argument for VAT—or, at least, the one *I* find most powerful—is the argument from *deference*. Wall gives the argument from deference its clearest formulation. Wall writes:

> Suppose that you are wise and that you have an excellent understanding of what is good for me. You know my talents, temperament and vulnerabilities and you know what types of projects would best suit my nature. Further suppose I know that you are a person of good will who cares about my well-being. Given these facts, we can ask: Would my life go better if I let you take control of it? Would it be a better life if I always turned to you for direction as to what I should do before I took up any project or commitment?
>
> Most of us strongly think the answers to these questions are 'no.' It can be reasonable to defer to the judgment of others some of the time in some circumstances; but a person who surrendered his or her judgment in all contexts would not lead a fully good human life.[20]

Wall asks us to consider turning over our lives to the deliberation and decisionmaking of another person who happens to be wise and benevolent, and hence would be trusted to make decisions that would garner substantial non-autonomy welfare benefits. He then asks whether it would be better for us to so turn over our lives, and answers 'no'. On this point, Wall offers a specific case:

> Person A is wise and has self-knowledge. He chooses projects that suit his nature and reflect his understanding of what is valuable and worthwhile. Person A leads a good, morally decent, life. Person B is also wise and has self-knowledge. But she finds the process of decision-making irksome and does not enjoy making important life decisions. Fortunately, Person B has a friend with the requisite wisdom, knowledge and good will to make these decisions for her. Person B lets this friend take over her affairs and she leads a good, morally decent, life.

> In this second case it can be asked whether Person A and Person B lead equally good lives? Once again, most of us strongly think 'no.'[21]

If we believe that Person B lives a worse life than Person A, even if we assume that Person B's friend has all the requisite 'wisdom, knowledge, and good will' to render Person B's life as good in terms of non-autonomy value as Person A's, then it would appear (or so the argument goes) that autonomy is intrinsically valuable.

3. The Autonomy Fallacy Introduced

Some might be tempted to resist the force of the argument from deference. For instance, Mikhail Valdman claims that it is no prudential burden to 'outsource self-government' or to engage in the deferential strategy employed by Person B. Valdman offers two main arguments for this claim. First, Valdman suggests that there is no intrinsic prudential burden to be found in turning small decisions, such as financial planning, over to an advisor.[22] And if this is correct, 'this introduces a puzzle, for if it were at least slightly bad in some respect to let another make any decision for you, then we could easily explain why it would be bad to let another make them all. But if it needn't be bad in any respect to outsource some of your decisions, why must it be bad in some respect to outsource all of them?'[23] In addition, Valdman argues that one could imagine a person whose 'decision mechanism' is, in a series of steps, removed from her brain. According to Valdman no step in this sequence renders this individual any worse-off, and so it's unclear why we should say that she is worse-off when, ultimately, her decision mechanism is controlled by a force entirely external to her.[24]

Valdman may be right. But for my money, I find the argument from deference powerful, and will accept for the purposes of argument that Person B lives a worse life than Person A. But, or so I claim, this has nothing to say in favor of the intrinsic value of autonomy. Note that there are many ways to distinguish the lives of Persons A and B. Recall that autonomy is not *simply* self-control over the activities and decisions of one's life. It is instead controlling one's life *in a certain way*; autonomous choice is a subset of the kind of choice one might make concerning how to run one's life. But Person B doesn't simply lack *autonomous* self-control of her life, she also lacks an important form of self-control I refer to here as 'self-direction *tout court*'.

The intuitive idea of self-direction *tout court* is of a person's own choice (whether a result of evaluative attitudes or bare first-order motivation), free

of external control, manipulation, or coercion. To state the idea somewhat more precisely, for a choice to be self-directed *tout court* it must meet both a positive and negative condition. The positive condition: a self-directed choice is made on the basis of, or *from*, an individual's pro-attitudes (whether first-order or higher-order). The negative condition: a self-directed choice must not be *mediated* by the agency or decisionmaking of others. If I ϕ because I want to, my ϕ-ing is self-directed. If I ϕ because the Pope told me to, or insisted that I should, or because controlled or manipulated me into ϕ-ing, this choice is not self-directed. Though I may have *some* pro-attitude directing me to ϕ (say, an interest in doing whatever the Pope says), this choice does not meet the second (negative) condition: it is not a choice made free of external direction or mediation. This account entails that some self-directed choice is made on the basis of whim or bare first-order desire and does not require autonomous control (whether deliberative or non-deliberative). (Note: I have stated the idea of self-direction *tout court* in terms of choices, but this could be put in terms of lives *mutatis mutandis*.)

Though there is a difference between Person A and Person B in respect of autonomy, there is also a difference between Person A and Person B with respect to self-direction *tout court*. Person B's decisionmaking is entirely mediated by the external control of her friend. Given that this is the case, Person B's life decisions (where to attend college, with whom to form romantic relationships, where to live, which restaurant to try, etc.) are not self-directed in the sense outlined here.[25]

I have not yet shown that we should reject the argument from deference. I have merely pointed out that, in principle anyway, there are differences between the lives of Person A and Person B that are not simply drawn along the dimension of autonomous self-government. This does not yet show that the argument from deference does not support VAT. At least one further thing must be shown, i.e., that self-direction *tout court* is itself valuable, or could explain why a self-directed life is or can be better than a non-self-directed life. If self-direction *tout court* is not *of itself* a feature of a life well-lived, the argument from deference does, in fact, support VAT. The remainder of this section, however, will argue for the following two claims. First, that self-direction *tout court* can plausibly help to explain the value of very important welfare goods (long-term 'projects'); second, that the distinctive evaluative contribution of such long-term projects is possessed by projects that are the product of *tout court* self-direction (even if non-autonomous). If these claims are true, the argument from deference cannot support the intrinsic value of autonomy. It commits the *Autonomy Fallacy*: a confusion

of autonomous self-control with self-direction *tout court*.

3.1. Self-Direction and its Value

There is good reason to believe that self-direction *tout court*, of the sort lacked by Person B, is important to the maintenance of a range of crucially important welfare goods.[26] I fix on one such reason here that I find extraordinarily plausible. But there may be others and I don't wish to rule them out.[27]

One thing worth noting, however, is that I shall not claim—and do not believe—that self-direction *tout court* is itself intrinsically good. I make only a somewhat weaker claim, i.e., that self-direction is an important *feature*, or *necessary condition* for other very important welfare goods. I leave it open that others may argue in favor of self-direction as *itself* an intrinsic value. (Either claim—or so I shall argue—can explain the disvalue of deference.) To see the importance of self-direction, take the following case:

> *Randall*: Randall is Native-American, born into a small tribe that continues to live on reservation land. His tribe has succumbed to serious social problems including rampant poverty and lack of education. Randall sets for himself the goal of improving the educational system for his tribe, works for his life to do so, and succeeds.

Randall, we might say, has a 'project': improving the educational system for his tribe, at which he works consistently throughout his life, and at which he succeeds. I think we should say that Randall's success at this project—that of improving the educational system on tribal lands—is an important factor in his overall welfare.[28] And its value goes beyond *simply* being valued by him and goes beyond the pleasure, or other individual momentary benefits, that success and pursuit of this project might bring. There are a number of common rationales for the welfare value of long-term projects of this kind. For instance, Velleman claims that such projects will help to shape an individual's life story and that in so doing, they provide Randall's life with a quality that cannot be explained *simply* by aggregating the welfare value of the individual moments in Randall's life.[29] Bernard Williams holds that some projects of this kind (in particular, the *ground projects*) will help to shape the *meaning* of Randall's life.[30] This fact, one might think, helps to explain the significance of Randall's project for the overall quality of his life; *just this sort of thing* constitutes whatever meaning Randall's life has. Both these explanations sound plausible to me, but whatever the rationale,

it seems right to say that this sort of achievement is intrinsically good in a special way.

But now consider Randall's success at this particular project in contrast to another substantial feature of his life, in particular, his *race*. We would not be tempted to say that the mere fact that he is a Native-American is an intrinsic contributing factor (negative or positive) to his well-being. But why? Importantly, neither Williams' nor Velleman's rationales for the special value of long-term projects can distinguish Randall's project from his race. Surely his race or nationality also contributes to the meaning of Randall's life; contributes, to substantial degree, to his life story and its overall 'shape'. But if this is correct, if the fact that projects like Randall's contribute to a person's life story and meaning is an important explanation of their value, it cannot be a sufficient explanation. There must be something else. I propose that the explanation is provided by the fact that Randall's project, and not his race, shapes his life story, and provides his life with a meaning, in a *self-directed way*. Long-term projects provide my life with a meaning that *I shape*; my born-into nationality or race does not. And if this is correct, it is plausible to say that a central welfare good, i.e., projects of the sort that Randall successfully engages, depend for their welfare value (or, at least, for their *distinctive* welfare value) on the fact that they are *self-directed*. Without being self-directed they would not, quite literally speaking, be *projects*.

3.2. Self-Direction versus Autonomous Self-Direction

Of course, it may be claimed that the argument so far does not establish that self-direction *tout court* is evaluatively significant. We might imagine, for instance, that Randall is perfectly autonomous. And hence the claim that Randall's project—rather than, say, his *race*—is valuable does not support the evaluative significance of self-direction *tout court rather than* the evaluative significance of autonomy. The partisan of VAT might simply respond that the evaluatively significant distinction between Randall's project and his race is that the former is chosen autonomously. The latter is not.

However, it would be a mistake to say that the distinctive welfare value of Randall's project must depend on whether it was the product of autonomous choice. Take, for instance:

> *Roger*: Roger is Native-American, born into a small tribe that continues to live on reservation land. His tribe has succumbed to serious social problems including rampant poverty and lack of

> education. Roger possesses a first-order motivation to improve the educational system on his tribe's land, and simply acts on it. (Perhaps his desire to improve conditions on his tribe's land is a result of unmediated *anger* or *sadness*.) He follows this motivation, works his life to improve the educational system of his tribe, and succeeds.

Roger chooses his project not on the basis of a higher-order evaluate state, but rather, simply on the basis of a first-order desire or motivation. Indeed, we may even imagine that Roger has a *disposition* to value this project, but simply fails to engage this disposition or act on the basis of this evaluative stand. (Imagine that though Roger never considers whether improving the educational system of his tribe conforms to his conception of value or reasons, *were he to have considered it*, he would have regarded this project as worthwhile, indeed. This project is reflected in Roger's conception of the good, but his decisions are never made *on that basis*. They are made, rather, on the basis of momentary whim. They simply *add up* to a self-directed project of great significance.) However, as a matter of considered judgment, the mere fact that Roger engaged in this project as a product of a first-order, rather than higher-order, pro-attitude does not entail that this project lacks the distinctive welfare value maintained by long-term projects and successful achievements. It would certainly be philosophically churlish to deny that Roger's life maintains the distinctive welfare value of long-term projects *simply* because he failed to exercise his capacity for autonomous control (whether deliberative or non-deliberative).

However, were it the case that Roger engaged his project because he was told to by a beneficent advisor, rather than strictly on the basis of his own motivation, we would appear less likely to treat this project as a significant welfare good. Indeed, it seems right to say that under such conditions, this project is less something that contributes to the distinctive value of Roger's life, but is instead something that—while significant—simply *happened to him*. Thus, it would appear that self-direction *tout court*, in the sense I mean here, is significant for the welfare value of such projects. And hence, if this is the case, we have reason to object to deference without relying on the intrinsic value of autonomy. Person B's life, after all, is *not self-directed*: all decisionmaking for Person B is made by the trusted advisor in question. Given this, whatever 'life story' or 'meaning' her life maintains should not be distinguished from the sort of story or meaning provided by other things over which she has no control, including her born-into gender, race, social class, etc. Person B's life will not maintain the distinctive welfare value

provided by long-term self-directed projects. And if self-direction *tout court* is sufficient to explain the distinctive welfare value of the sort of projects on display in Randall's cases (as I have so far argued that it is), the lack of self-direction *tout court*, rather than the lack of *autonomous* self-direction, is sufficient to explain why Person B lives a worse life that Person A. And, hence, the argument from deference commits the Autonomy Fallacy.

3.3. Response: Is All Self-Direction Autonomous?

One might object to my argument in the following way. I've so far been holding that self-direction *tout court* should be distinguished from autonomous choice. But it might be argued that *without* autonomy, one's life *cannot* be self-directed. For instance, Robert Young writes:

> [A]utonomy is part of the moral basis of personhood; it transforms what would otherwise be utterly episodic (and hence not *the life of a person*). To the extent that a person is at the mercy of his (or her) urges or impulses, or lacks scope for actively planning and realising goals and purposes, it is the person's circumstances, not the person himself (or herself), that governs. Accordingly, the person's life will lack self-direction.[31]

Young's thought appears to be this. If we assume that Roger's 'urges and impulses' (i.e., first-order motivations) govern the direction of his life, Roger's life is controlled by 'circumstances', rather than by his own self-direction. For Roger's life to be autonomous it must be the case that Roger controls his impulses *via* some sort of second-order procedure (whether cognitive, etc.). And hence all self-direction is necessarily autonomous self-direction, and hence there is no Autonomy Fallacy.

This argument is not plausible. There's surely a distinction between a life being dictated by circumstances and being chosen on the basis of a bare first-order motivation. It seems right to say, for instance, that someone chased by a tiger, or someone who is under the thumb of dictatorial oppression, lives a life dictated by circumstances.[32] But contrast this life with, e.g., a rich gadabout, who simply follows his own unreflective impulses ('drifts though life') without engaging his higher-order conception of the good. I find it very difficult to say that there is no difference in terms of self-direction between the former and latter persons. The former lack an important concept of self-direction; their choices are dictated by external forces. But there is certainly an important distinction worth drawing between this person and the second. Of course, it is certainly true that a person who is chased by

a tiger is fulfulling a substantial degree of her first-order motivations, i.e., to keep away from the tiger. But this does not entail that her life is not dictated to her by external forces.[33]

Thus we can surely admit that there is at least *some* distinction (perhaps, admittedly, with a healthy amount of gray area) to be drawn between (a) a person whose life is dictated by others (such as a dictatorial oppressor), and (b) someone whose life is lived according to his or her own desires, but who lacks the relevant conditions of *autonomous* choice. Furthermore, this difference matters. There is a very substantial evaluative difference, or so it seems to me, between the impact on one's welfare of choices that are made under external duress or oppression and simply on the basis of first-order whim. This is illustrated clearly by the argument from deference and by a consideration of Roger. If we wish to say that Roger's long-term project of improving conditions on his tribe's land maintains the *distinctive* welfare value of long-term projects and achievements (as I think we should), it would appear that the second style of choice maintains an evaluative character that is missing from a life that is solely determined by, e.g., crushing poverty, being chased by a tiger, or having one's decisions made by a beneficent advisor, etc. And this is all that is required to show that the argument from deference succumbs to the Autonomy Fallacy: one can explain the evaluative distinction between the lives of Persons A and B *without* reference to the intrinsic value of the particular style of decisionmaking deemed *autonomous* (however one understands the nature of autonomous choice).

4. Does Autonomy Add Value?

So far the dialectic of this paper runs as follows. A significant argument in favor of VAT is the argument from deference. But, or so I claim, the argument from deference commits The Autonomy Fallacy: the disvalue of deference can be explained by self-direction *tout court*, which we have independent reason to believe is evaluatively significant, rather than autonomous self-direction. And so the argument from deference cannot support VAT.

But so what? All I've shown is that there is a gap in the argument from deference. Nothing in the argument I have so far offered, however, tells against any other potential arguments for VAT. But, or so I hope to show, once we are cognizant of the Autonomy Fallacy (as illustrated by the flaws in the argument from deference), the claim that autonomy *of itself* is an intrinsic benefit (in comparison to the evaluative significance of self-direction *tout court*) becomes much less plausible. If VAT is to hold, the fact of autonomous deliberation must *add value* to choices that are otherwise

self-directed. In what remains I consider three interpretations under which one might accept the claim that autonomous choice adds value in this way. I claim that none are plausible. And hence it would appear that VAT is only plausible if *we* commit the Autonomy Fallacy, i.e., confuse autonomous self-government with self-direction *tout court*.

As I noted before, I will focus here on the autonomy of individual choices and their contribution to the value of a life. I argue, in §4.4, that this focus also sheds substantial doubt on views who would insist on the intrinsic value of autonomous lives.

4.1. The Natural Interpretation

One reason to believe that autonomy adds value to self-direction *tout court* is simply that autonomous choices are intrinsically valuable in comparison to other sorts of choices one might make about one's life. Call this the 'Natural Interpretation' of the additive value of autonomy. However, the Natural Interpretation faces a challenge. Some autonomous choice can be—as I shall use the term—*unsuccessful*: one might decide to ϕ on grounds of a commitment to the value of p, which ϕ-ing is intended to promote, but then come to realize that one's own assessment of p's value is unstable or incoherent with other things one values; one might deliberate and select ϕ on the basis of reasons one recognizes in favor of ϕ-ing, but then come to recognize that the reasons one acted on were not genuine reasons, or that these reasons actually tell in favor of some alternative action or state of affairs, etc. Alternatively, one might simply fail to achieve that which one autonomously decided to pursue. For instance, one might autonomously decide to become a lawyer, but fail to be hired by a law firm. More simply, one might choose to ϕ at t_1 on the basis that ϕ-ing will produce some sought-after good at t_3, but then drop dead at t_2. This does not entail that an autonomous choice has not been made. Rather, it simply means that this particular autonomous choice was unsuccessful.

The Natural Interpretation would appear to hold that *unsuccessful* autonomy can improve a person's life in comparison to a choice that is otherwise self-directed *tout court*. But I find it difficult to believe that we would hold that unsuccessful autonomous self-direction *itself* adds value to the mere fact of self-direction.[34] Here's an example that, I think, confirms this general thought:

> *Unsuccessful Autonomy: Madeline and Gussy*: Madeline and Gussy have been spending a lovely afternoon together walking

> in country gardens and discussing mutual interests. Gussy is very attracted to Madeline and *vice versa*, though neither of them really knows quite what to do about it or how to proceed. At some point, they both plan and choose to kiss on the basis of their autonomous reflection, on the basis of reasons they regard as telling in favor of the kiss. They both believe that such a kiss will be romantic, and will be something they will cherish as a valued memory, even if nothing comes of their relationship. But this kiss, as it turns out, is utterly unromantic; an expression of their mutual awkwardness rather than any genuine romance. Neither of them comes to value the kiss or remember it with affection.

Here we have an example of an autonomous choice, in particular, the autonomous choice to *kiss*. This kiss, had it been romantic or valued by either of them, would surely have been intrinsically valuable. But as it happens, their kiss was entirely unromantic and unvalued. But should we say that, despite the fact that the kiss fails to maintain any of the sought-after value, their autonomous choice of itself improves their lives?

It's important, in answering this question, to have a proper contrast in mind. Now let's consider a slightly different version of Madeline and Gussy:

> *Unsuccessful Self-Direction: Madeline and Gussy*: Madeline and Gussy have been spending a lovely afternoon together walking in country gardens and discussing mutual interests. Gussy is very attracted to Madeline and *vice versa*, though neither of them really knows quite what to do about it or how to proceed. In the grip of passion, Gussy kisses Madeline. (Gussy would later tell a friend that this impulse 'took hold of him'.) But this kiss, as it turns out, is utterly unromantic; an expression of their mutual awkwardness rather than any genuine romance. Neither of them comes to value the kiss or remember it with affection.

In this case, the result is the same as the autonomous kiss: it lacks romance, and it is valued by neither Gussy nor Madeline. According to the Natural Interpretation, Madeline and Gussy's autonomous kiss renders their lives *better for them* than the kiss that is not the product of autonomous choice. But this proposal is very difficult to believe. This answer, in large measure, can be explained by the fact that neither Gussy nor Madeline's autonomous deliberation resulted in anything of genuine value. Unsuccessful autonomous choice does not add value to a life in comparison to self-direction *tout court*.

And hence the Natural Interpretation of the additive value of autonomy is false.[35]

One might object that, in this case, I have focused on the stronger, deliberative, account of autonomous control. But my focus here is legitimate. Note that in the first version of their kiss, Madeline and Gussy display *all* relevant notions of control. And hence if autonomy adds value on either account, one would expect their kiss to be better in the first case than in the second. After all, it is hard to see how the possession of a much *weaker* form of autonomous control (e.g., as suggested by Wall) could improve a life when the possession of a much *stronger* form of control does not. Here's another way to put this point. In the original case, Madeline and Gussie possess both deliberative and non-deliberative notions of autonomous control. But their unsuccessful autonomous kiss is no better for being autonomous. To claim that autonomy still retains value in the face of this judgment, one would have to claim that the mere presence of deliberative control *cancels* or somehow *voids* the value of their possession of non-deliberative control. But why should this be so? Even if we don't wish to say that individuals *must* deliberate to obtain the intrinsic value of autonomy, why downgrade the extent to which autonomy is valuable if they do?[36] This proposal seems too strange to take seriously. And hence I conclude that unsuccessful autonomy (whether deliberative or non-deliberative) does not add value to unsuccessful self-directed choice.

4.2. The Success Interpretation

The Natural Interpretation fails because it seems implausible to say that unsuccessful autonomy should be of any added value over unsuccessful self-direction. But one can interpret the additive value of autonomy differently. In contrast to the Natural Interpretation, the 'Success Interpretation' would value only *successful* autonomy.

Indeed, a schematic representation of this proposal is suggested by Wall. Wall claims that though autonomy is intrinsically valuable, it is not *unconditionally* valuable. That is, though it is valuable for itself, and not for the sake of anything else, its intrinsic value is conditional on other things, including (potentially) success. Another way to put this suggestion is that the intrinsic value of autonomy does not supervene on the intrinsic properties of any particular choice (i.e., which would yield the unavoidable claim that unsuccessful autonomy is just as valuable as successful autonomy). Rather, the intrinsic value of autonomy depends on further, extrinsic properties (such as success).[37] Under such conditions, autonomy adds value to self-direction

tout court.

I think there is good reason to reject this view. Consider another version of Madeline and Gussy:

> *Successful Autonomy: Madeline and Gussy*: Madeline and Gussy have been spending a lovely afternoon together walking in country gardens and discussing mutual interests. Gussy is very attracted to Madeline and *vice versa*, though neither of them really knows quite what to do about it or how to proceed. At some point, they both choose to kiss on the basis of their autonomous reflection, on the basis of reasons they regard as telling in favor of the kiss (after due deliberation). They both believe that such a kiss will be romantic, and will be something they will cherish as a valued memory, even if nothing comes of their relationship. They have an enjoyable romantic kiss, which they both come to value.

Now consider another version:

> *Successful Self-Direction: Madeline and Gussy*: Madeline and Gussy have been spending a lovely afternoon together walking in country gardens and discussing mutual interests. Gussy is very attracted to Madeline and *vice versa*, though neither of them really knows quite what to do about it or how to proceed. At one point Madeline kisses Gussy, simply as a result of a bare impulse. (Madeline would later tell a friend that this impulse 'took hold of her'.) They both regard this kiss as enjoyable and romantic, and come to value it.

In the second case, neither Madeline nor Gussy exercise their capacities for autonomous choice in kissing the other. Rather, the kiss is the product of a single, unreflective desire rather than any higher-order evaluative state (such as a belief or conception of the good, etc.). But should we think that Madeline and Gussy's lives are better, more choiceworthy, if they kiss as a result of their autonomous deliberation or as a product of a higher-order evaluative attitude, rather than kissing simply because they want to, or as a result of being in the grip of a passionate drive or first-order desire? This proposal strikes me as utterly, egregiously absurd.

4.3. The Significance Interpretation

The partisan of VAT might be, at this point, impatient. In discussing the value of self-direction *tout court*, I noted that self-direction *tout court* is an important aspect of a range of welfare goods, i.e., long term *projects* or pursuits that characterize an important aspect of a person's life story. But one proposal might be that autonomy does not add value to *every* instance of self-direction. Rather, only autonomous choices that lead to long term projects or pursuits add value to self-direction *tout court*. Call this 'The Significance Interpretation'.

Indeed, there may be some rationale for this suggestion. Recall Valdman's first argument against the appeal to deference. Valdman holds that there can be no evaluatively significant distinction between outsourcing one's decisions about one's own life and outsourcing one's, say, financial decisions to a trusted advisor. But if we allow that there is an important distinction between choices one makes that do, and do not, affect the long-term structure of a life, there is a principled reason for fans of VAT to resist the claim that turning over one's financial decisions makes one worse-off, but to hold fast to the claim that turning over *all* of one's decisions does make you worse-off. And so there may be good reason to believe that autonomy adds value *only* to the sort of choice that characterizes the long-term structure, meaning, or character of a life.

I think there are two problems with VAT on this interpretation. First, there is no reason to believe that momentary goods like the kiss between Madeline and Gussy couldn't affect the long-term meaning and character of their lives. In particular, this kiss might have been a turning point, a crucial moment in their lives' stories. It may go on to shape the long-term structure of their lives in important ways. But even if it is such a turning point, we are no more likely to say that it would be better for them had they chosen to kiss autonomously rather than non-autonomously.

Second, it is implausible to believe that autonomy, even when shaping an individual's longer term projects, adds value to self-direction *tout court*. Two cases shed light:

> *Significant Autonomy: Tuppy and Angela*: Tuppy has recently had a spat with his longtime fiancé Angela. Tuppy recognizes that his own decisionmaking is clouded by his anger at Angela's criticism of his weight, and her refusal to accept his apology for minimizing a recent traumatic event in her life. Given his spat, Tuppy is very seriously thinking of marrying another woman, Cora. After much autonomous deliberation, Tuppy decides to set aside his engagement to Angela on the basis of his recognition

> of the comparative value of marrying Cora. Tuppy and Cora maintain a long and happy marriage, but one marked by less substantial happiness than Tuppy's marriage to Angela would have been.

Compare this case to:

> *Significant Self-Direction: Tuppy and Angela*: Tuppy has recently had a spat with his longtime fiancé Angela. Tuppy recognizes that his own decisionmaking is clouded by his anger at Angela's continuing to criticize his weight, and her refusal to accept his apology for minimizing a recent traumatic event in her life. Overcome by his anger at Angela's intransigence, and allowing himself to give in to his passion for Cora, he breaks his engagement to Angela and marries Cora. Tuppy and Cora maintain a long and happy marriage, but one marked by less substantial happiness than Tuppy's marriage to Angela would have been.

On the current proposal, autonomy adds value to self-direction *tout court* to the extent that it is efficacious at shaping the long term structure and meaning of one's life. On this view, Tuppy's exercise of autonomy in deciding to marry Cora is intrinsically valuable. After all, this marriage shapes the long-term meaning of his life and life story; this instance of (deliberative and non-deliberative) autonomous choice surely passes the relevant threshold of significance. Hence deciding to break his engagement to Angela autonomously rather than as a result of his own whim renders his life better, despite the fact that, in both cases, the marriages are equally sub-optimal in comparison to his potential marriage to Angela. This, I claim, is implausible. When we note that in *both* cases the marriage was self-directed, it becomes much less plausible to believe that allowing this first-order motivation to control his choices rather than his capacity for autonomous deliberation would *in itself* make his life worse. Autonomy, whether deliberative or non-deliberative, does not add value to self-direction *tout court*.

4.4. Choices and Lives[38]

So far the arguments I present in this section have shown (or so I boldly claim) that autonomous choices do not add value in comparison to choices that are self-directed *tout court*. This seems to show that autonomous (rather than self-directed) choices should not bear intrinsic value. Of course,

there is another interpretation of VAT: that autonomous *lives*, rather than choices, bear intrinsic value. But (or so I shall now argue), the very same cases show that we should reject the claim that autonomous *lives* are intrinsically valuable (at least in comparison to lives that are relevantly self-directed).

Two claims seem to me to generate this result. First, a descriptive claim: given that the extent to which a life is autonomous supervenes on the extent to which the choices in that life are autonomous, we should say that, other things equal, a life gets *more autonomous* when there are *more choices* in that life that are autonomous.[39] This seems straightforward enough. Second, an axiological claim. If autonomous lives are intrinsically valuable, we should say that, other things equal, the more autonomous a life gets, the more intrinsically valuable it is.[40] But the cases on display in §§4.1-4.3 seem to indicate that lives that are more autonomous, given the addition of autonomous choices, are not thereby better in comparison to lives that are more self-directed. And hence we should deny that autonomous lives, any more than autonomous choices, bear intrinsic value as endorsed by VAT.

You could reject this line of reasoning. First, you might deny the axiological claim, holding instead that a life does not get better, other things equal, when it gets more autonomous. Instead, one could hold that lives get better (other things equal) to the extent that they maintain some sufficient *threshold* of autonomy. This would imply that lives do not get better the more autonomous they are; they only get better when they achieve this relevant threshold or 'mark' of sufficient autonomy in a life. But this proposal faces two challenges. First, it is implausible. If we are willing to commit to the intrinsic value of autonomy (whether in a life or choice) as insisted upon by VAT, we should also hold that *the more autonomy the better* (whether the autonomous choice or autonomous life is bearer of this intrinsic value). The proposal on offer would seem to say that the only time an autonomous choice makes a difference to the quality of life is when this autonomous choice spells the difference between an autonomous life and a non-autonomous one. All other improvements in the autonomy of a life are of no value whatsoever. Though not incoherent, I find this implausible.

Second, even if we deny that a life gets better (other things equal) to the extent that it gets more autonomous, this is no more plausible *given* the cases I outline above. In particular, we could easily construe the exercises of autonomy in any of the cases above as marking the difference between a life of insufficient autonomy and one of sufficient autonomy. But, or so I claim, it is no more plausible to say that, e.g., Madeline's successful exercise of autonomy in kissing Gussie renders her life better in comparison to her

passionate kiss *even if* there is some threshold involved. Here's another way to put this point. To know whether autonomy adds value for Gussie, Madeline, or Tuppy, we would have to know the extent to which they have or have not exercised autonomy in the past, or the extent to which this choice would yield the relevant 'threshold' of autonomy, whatever it is. Given the cases as stated (without this information), we should reserve judgment. But this isn't right: we don't reserve judgment nor do we feel we should. Whatever the extent to which their lives are autonomous, it seems implausible to say that *this choice*, made autonomously, adds value to their lives in comparison to the same choice, self-directed. Thus even if we were to deny that autonomy adds value in a scalar fashion, this is no help for the partisan of the intrinsic value of autonomous lives.

Alternatively, one could deny the principle that, other things equal, a life gets more autonomous to the extent that more of the the choices in said life are autonomous. Of course, this principle might be true. But it is unhelpful in the present argument even if true *unless* Gussie, Madeline, and Tuppy's autonomous choices themselves fail to contribute to the autonomy of their lives. But this is especially implausible in Tuppy's case, as the choice of an individual's long-term romantic partner is surely significant enough to contribute to the extent to which his life is autonomous. And it also seems implausible in Gussie and Madeline's cases. Imagine that though their lives are otherwise identical, Gussy kisses Madeline autonomously, Madeline kisses Gussy as a product of non-autonomous self-direction. Though it is a relatively small thing, Gussy's life was to that extent more autonomous than Madeline's. To say otherwise would appear to commit one to denying that small, momentary exercises of autonomy cannot contribute to the overall autonomy of a life. But that's not plausible: just as short-term mild pleasures can contribute to the overall pleasurableness of a life, small, momentary exercises of autonomy surely contribute to the overall autonomy of a life. I therefore conclude that if the autonomy of Gussy, Madeline, and Tuppy's choices does not contribute to the value of their lives, we should hold that the autonomy of a life, as well as the autonomy of choices, fails to add value in comparison to self-direction *tout court*.

4.5. The Autonomy Fallacy Revisited

The Autonomy Fallacy is displayed in at least one central argument for the welfare value of autonomy: the argument from deference. There is no reason to believe that the prudential burden of deference *must* be explained with reference to the intrinsic value of autonomy. But the argument in this

section seems to show that the Autonomy Fallacy is more widespread. Once we distinguish *autonomous* self-direction from self-direction *tout court*, the claim that autonomy itself is intrinsically valuable is much less tenable as a substantive theory of well-being (whether or not self-direction *tout court* is intrinsically valuable or otherwise evaluatively significant). And so there is good reason to doubt VAT: VAT is most plausible when we confuse autonomous self-government with the mere lack of external compulsion or control, i.e., when *we* commit the Autonomy Fallacy.

Of course, there may remain a bare form of skepticism about the denial of VAT. Some might simply believe that a person's life cannot go well, or cannot go well to a particular degree, if that life is not autonomous to at least the barest extent. But this skepticism, stated as such, seems to me to display precisely the fallacy I'm interested in exposing here. Take, for instance, Roger. Roger's life is incredibly successful; he acts on his desire to rid his tribal area of poverty, and succeeds at so doing. Roger's life is extremely good, especially so *given* that we can simply stipulate that Roger has the *disposition* (though he does not act on it) to value the activities in which he is engaged. Of course, this is no more than a bare intuition; but once the distinction between autonomous and non-autonomous self-direction is made clear, skepticism about the denial of VAT appears much less plausible.[41]

Of course, this more general conclusion relies upon substantive considered judgments that not all will share. Nevertheless, even if my analysis in the cases provided are not responsive to the reader's judgments, I hope to have shown, at least, that the most significant rationale for VAT in response to an important form of skepticism displays the Autonomy Fallacy and should thereby be rejected. It is still possible that we may find VAT plausible when all is said and done. But I hope to have provided at least some *prima facie* reason for being skeptical of that ultimate outcome.

5. Conclusion

In this paper, I have argued as follows. I begun by considering one particular argument for the intrinsic value of autonomy: the argument from deference. I then showed that this argument displays the 'Autonomy Fallacy': a confusion between autonomy and self-direction *tout court*, or the mere lack of external control over one's life and activities. Once we recognize that self-direction *tout court* is itself evaluatively significant, we can fully explain the disvalue of deference (of the kind undertaken by Person B) without reference to the intrinsic value of autonomy.

But I then argued that a recognition of the Autonomy Fallacy creates

problems for VAT on the wholesale. Once we clearly distinguish between self-direction *tout court* and *autonomous* self-direction, the claim that autonomous decisionmaking *in particular* improves the value of a life becomes much less plausible. I considered three potential interpretations of the additive value of autonomy. None can plausibly deliver the additive value of autonomy in comparison to a life of self-direction *tout court* (whether or not self-direction *tout court* has evaluative significance, which it may well lack in some cases).

Some may be concerned that the rejection of VAT will have very substantial implications for the attractive moral and political conclusions that seem (in at least some arguments) to rely on something like VAT. But this concern is, in large measure, unwarranted. The evaluative significance of self-direction *tout court* is or should be substantial enough to establish many of the *per se* moral or political conclusions that are often made on the basis of the value of autonomy. For instance, if we believe that self-direction *tout court* is valuable, we continue to have grounds to object to some forms of state molestation or legislation that would direct people to live one way rather than another. Insofar as such legislation would be an example of *external* control (rather than simply non-autonomous control), the evaluative significance of self-direction can continue to play a role in political argument. Whether the evaluative significance of self-direction *tout court* can play *all* the roles in moral or political argument that have previously been played by the intrinsic value of autonomy is a further question I won't broach here. But there is very good reason to believe that any conclusions that can only be supported by accepting the intrinsic welfare value of autonomy are unsupported.[42]

Department of Philosophy
University of Kansas

Notes

[1]Cf. Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1988), 390.

[2]By 'subjectivist' I mean the class of views for which a necessary condition of ϕ's value for x is x's (suitably construed) pro-attitude toward ϕ. See Dale Dorsey, "Subjectivism without Desire" in *The Philosophical Review* 121 (2012).

[3]Wall, 129-30.

[4]George Sher, *Beyond Neutrality: Perfectionism and Politics* (Cambridge: Cambridge University Press, 1997), 176.

[5]Kymlicka, *Liberalism, Community, and Culture* (Oxford: Oxford University Press, 1989), 12.

[6]John Stuart Mill, *On Liberty* III.4.

[7]David Brink, "The Significance of Desire" in *Oxford Studies in Metaethics* v. 3, ed. Shafer-Landau (Oxford: Oxford University Press, 2008), 31-45.

[8]Kymlicka, 12-13.

[9]Wall, 206.

[10]Mill uses this argument to reject censorship, among other liberty-interfering forms of state legislation. See *OL* II, 6.

[11]Nomy Arpaly, *Unprincipled Virtue* (Oxford: Oxford University Press, 2003), 118.

[12]Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge: Cambridge University Press, 1988), 20.

[13]Brink, 41.

[14]Sher, 47.

[15]Thomas Hurka, *Perfectionism* (Oxford: Oxford University Press, 1993), 151.

[16]Wall, 139.

[17]I should note that I use 'higher-order' here to identify any particular attitude that is not simply a first-order motivation. Second-order desire will count, for instance. But so will evaluative *beliefs*: beliefs that something is worth doing or is good, say. (As will beliefs that evaluate or endorse the objects of first-order motivations.) I intend this term to be broadly ecumenical.

[18]Wall, 136-9.

[19]Sher, 48.

[20]Wall, 146.

[21]Wall, 147.

[22]Mikhail Valdman, "Outsourcing Self-Government" in *Ethics* 120 (2010).

[23]Valdman, 772.

[24]Valdman, 777.

[25]An anonymous reviewer challenges this. There could be two aspects of Person B's life that are self-directed. First, it may be that Person B has an antecedent desire to ϕ, but decides to ϕ *because* her friend told her to do so. And hence, Person B's life may be entirely self-directed, failing to establish a distinction between A and B. However, this objection relies on a faulty understanding of self-direction *tout court*. The mere fact that I desired ϕ prior to being told do ϕ does not entail that my ϕ-ing is self-directed. It must be the case that *having ϕ'd* was a product of this first-order motivation *in the absence* of external coercion or control, which is clearly not present in the case of Person B. Put bluntly, even if Person B maintained an antecedent desire to ϕ, this does not entail that her choice to ϕ meets the negative condition of self-direction. Second, it could be that Person B has a perfectly self-directed motivation to *conform to whatever her friend says she should do* (*de dicto*, as it were). After all, her choice to conform to whatever her friend tells her to do meets both the positive *and* negative conditions for self-direction *tout court*. And if this is correct, the extent to which Person B's life is self-directed cannot illuminate a difference with Person A. I do not find this compelling. While it is clear that there are *some* choices Person B makes that are self-directed, the choices that play the largest role in Person B's life are not: what college to go to, what sort of a person to be, with whom to engage in romatic relationships, etc. These choices fail the negative condition of self-direction *tout court*. And while there may be *some* dimension of Person B's life that is self-directed, this does not establish that self-direction shows any *less* of a difference between their lives than the extent to which A and B are autonomous: after

all, Person B may *autonomously* decide to do whatever her friend decides she should do. Indeed, given Wall's description of Person B, it seems entirely plausible to say that Person B *has* decided, autonomously, to turn over her decisionmaking to her friend. But, for the same reason, this is not enough to say that autonomy can't make a distinction between A and B: the autonomy that really matters (according to the partisan of VAT) is the autonomy when it comes to the specific choice of projects, relationships, etc. And so *if* the autonomy of their lives can illuminate an evaluative difference between A and B, so can self-direction *tout court*.

[26]I explore this idea in much more detail in *The Basic Minimum: A Welfarist Approach* (Cambridge: Cambridge University Press, 2012), ch. 2.

[27]For instance, Raz argues that one reason that autonomy is valuable is that in societies like ours, with certain recognized social forms, the ideal of choice is treated as one important feature of the goods we might undertake, including, e.g., marriage. (Raz, 392.) And though Raz holds that this is an argument for the claim that '[a]utonomy is a distinct ideal,' (Raz, 395) it is in fact an argument for the claim that self-direction is a distinct ideal, rather than any one particular *method* of self-direction. (Raz's theory of autonomy goes far beyond a simple lack of external coercion; see Raz, 372-3. Hence it seems to me that Raz's argument, like the argument from deference, is an example of the Autonomy Fallacy.)

[28]The importance of such 'projects' to welfare is accepted by many, including Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1987), ch. 12; Simon Keller, "Welfare and the Achievement of Goals" in *Philosophical Studies* 116 (2004); T. M. Scanlon, *What We Owe To Each Other* (Cambridge, MA: Harvard University Press, 1998), ch. 3; John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971) §64; Douglas Portmore, "Welfare, Achievement, and Self-Sacrifice" in *Journal of Ethics and Social Philosophy* 3 (2007).

[29]See David Velleman, "Well-being and Time" in *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000), ch. 3.

[30]See Bernard Williams, "Persons, Character, and Morality" in *Moral Luck* (Cambridge: Cambridge University Press, 1981), 12.

[31]Robert Young, "The Value of Autonomy" in *The Philosophical Quarterly* 32 (1982), 38. I should note that Young himself regards autonomy as valuable *as a result* of a 'non-instrumental desire' (Young, 43). However, he provides a further argument to believe that autonomy is intrinsically valuable (or that, for instance, one should non-instrumentally desire autonomy). Young notes that Nozick's experience machine seems to indicate that autonomy, of itself, is valuable. But this argument commits the very fallacy I have so far been at pains to illustrate: in the experience machine, one is not simply non-autonomous, but one does not make choices about one's life *to any degree*, whether by autonomous control of first-order motivations, or by means of those first-order motivations themselves. One's life is simply dictated.

[32]The tiger example is infamous from Joseph Raz. See Raz, 374.

[33]See note 25.

[34]Would we hold that self-direction *tout court* is intrinsically valuable if it is unsuccessful? In a way this question is moot, insofar as the view I propose does not treat self-direction *itself* as intrinsically valuable; self-direction is, rather, a necessary condition of the distinctive value of self-directed projects. But no matter how one comes down on this question, it seems relatively clear that autonomous self-direction does not add any *additional* value if it is unsuccessful.

[35]Notice that nothing in this argument is committed to the claim that mere *self-direction*

is evaluatively significant here. What is essential for this argument is only that there is no recognizable evaluative distinction to be made between a life that exhibits unsuccessful self-direction and one that exhibits unsuccessful autonomous reflection, whether or not one regards the former concept as having any evaluative significance at all (which, it seems to me, is rather implausible).

[36]Notably, I'm not assuming that, in the first case, Gussie and Madeline's autonomous deliberation was instrumentally disvaluable (which may, perhaps, explain why the intrinsic value of a weaker form of autonomous control which is also displayed is cancelled out, say). I'm only assuming that Gussie and Madeline possessed every relevant form of autonomous control, but this autonomous control led to something disvaluable. But I do not stipulate that the kiss was disvaluable *as a direct result of* their autonomous deliberation.

[37]'[VAT] does not mean that personal autonomy is the only component of a fully good life. Far from it. By itself, and in isolation from other components, it has no value. Its value is dependent on the presence of these other components. Still, this dependence does not show that autonomy has no intrinsic value or that it is not a central component of a fully good life,' (Wall, 130). Note that there are two ways one might interpret Wall's claim that autonomy possesses only conditional value. First, one might hold that only successful autonomy is intrinsically valuable. Second, one might hold that autonomy is intrinsically valuable in cases in which other, non-autonomy goods are present in a life. But the latter interpretation seems implausible for reasons we have already given; it seems wrong to say that Gussy and Madeline's autonomous kiss is any improvement over their non-autonomous kiss, even if they both have a sufficient degree of non-autonomy goods throughout their lives.

[38]Thanks to an anonymous reviewer for suggesting this objection.

[39]The 'other things equal' clause is meant to hold fixed, e.g., the potential degrees of autonomy on display in a choice and the extent to which a particular choice may or may not be very significant concerning the shape of a person's life.

[40]This 'other things equal' clause is meant to hold fixed the other potential goods a life might obtain.

[41]Thanks to an anonymous reviewer for expressing this skepticism clearly.

[42]I'd like to thank the Murphy Institute, Tulane University, for generous support of this project. I would also like to thank an anonymous reviewer for PPQ for extraordinarily helpful comments.