Plan-Based Expressivism and Innocent Mistakes*

Steve Daskal

Published in 2009 in *Ethics* 119 (2):310-335
http://www.jstor.org/stable/10.1086/596458

Over the past hundred years, there has been a series of metaethical views according to which we ought to deepen our understanding of normative terms by inquiring after the states of mind they express. This is the class of views, sometimes identified as forms of non-cognitivism, that Allan Gibbard has dubbed "expressivist." It includes the emotivism of A. J. Ayer and Charles Stevenson, as well as, on some readings, the prescriptivism of R. M. Hare. More recently, Simon Blackburn's "quasi-realist" projectivism is a sophisticated form of expressivism, and Gibbard has also developed and defended an expressivist theory of his own.[1] What these views share is the thought that normative terms are intimately linked to emotions or actions in a way that cannot be captured in a purely descriptive analysis. They also share a set of common objections. For instance, all expressivists must confront the so-called "Frege-Geach" problem, according to which an expressivist analysis of normative terms will be unable to account for the many ways in which such terms can be embedded in otherwise descriptive sentences.[2] Expressivists also face objections, raised most prominently by Ronald Dworkin, according to which their analyses of normative terms fail to provide an adequate degree of objectivity for normative judgments.[3]

My own view is that expressivists such as Gibbard and Blackburn have been successful in responding to the objections involving embedded contexts, but I do not engage in that debate here. Instead, I focus on what I take to be the more compelling set of objections to expressivism, those related to concerns about objectivity. My primary aim is to raise a particular form of this objection in the context of the version of expressivism presented in Gibbard's book *Thinking How to Live*, which I take to be the most advanced form of expressivism to date.[4] As I see it, the real problem for Gibbard's latest form of expressivism is that there is a subset of our normative statements that cannot be subsumed under his expressivist analysis. This limit on the scope of his analysis leads me to identify a specific sense in which his view fails to capture the objectivity of normative claims.

I therefore begin, in Section I, with a brief summary of Gibbard's argument for expressivism. This sets the stage for the problem I intend to raise, which arises in the context of thinking about Gibbard's expressivist analysis of two forms of innocent mistakes.[5] In Section II, I consider the possibility of what I call pure planning innocent mistakes. This sort of mistake can occur in a situation in which the appropriate action differs from the action dictated by the appropriate plan, and this difference is not a result of the agent's poor epistemic situation. I say more about how to understand these pure planning innocent mistakes later, but for now let me note that I will argue, against Gibbard, that we ought not allow the possibility of such mistakes. This leads to Section III, in which I consider

another kind of innocent mistake, which I call an innocent mistake of mental constitution, that arises when an agent is constituted in such a way as to be incapable of making the correct decision. In considering what to say about this sort of mistake, I argue that such mistakes reveal a shortcoming of Gibbard's view by highlighting a subset of our normative judgments, and a corresponding sense of moral objectivity, that he is unable to capture. These judgments arise, or so I claim, in certain cases involving partial endorsement of an action. Gibbard might have hoped to capture this partial endorsement through the attribution of a pure planning innocent mistake, but that move is ruled out by the argument of Section II. Moreover, I further argue that other possible attempts to capture this partial endorsement in Gibbard's expressivist analysis, such as by appealing to his earlier norm-expressivist account of morality, are unsuccessful. Finally, in Section IV, I suggest that this problem is not peculiar to Gibbard's view but is instead likely to generalize to any form of expressivism that shares Gibbard's commitment to a fundamental connection between normative judgments and action.

I. GIBBARD'S PLAN-BASED EXPRESSIVISM

As indicated earlier, my objections to expressivism will be aimed most directly at the version of the view developed in Gibbard's book *Thinking How to Live*. This new version of expressivism is focused on the processes of reaching decisions and making plans. The basic idea is that questions involving "oughts," i.e., normative

questions, are essentially questions of what to do (or think or feel).[6] This leads Gibbard to posit a technical concept, expressed by the phrase "thing to do." The thought is that in deciding to perform a given action, one applies this concept "thing to do" to the action. To conclude that something is now the thing to do is therefore to express a decision to do it. More generally, to conclude that in any given situation a particular action is the thing to do is to express a plan to do it if in that situation. Gibbard simply stipulates that the concept "thing to do" works in this way, which is to say that he stipulates that an expressivistic reading of the concept is appropriate.

Gibbard's central hypothesis is that ordinary normative judgments actually are "thing to do" judgments. According to this hypothesis, normative claims are best understood as expressing planning judgments. So for instance, on this view, to conclude that right now because my throat is dry I ought to take a drink is to conclude that taking a drink is the thing to do in Gibbard's sense of the phrase "thing to do." If I say "Because my throat is dry I ought now to take a drink," I am expressing my decision to take a drink now. Similarly, if you were to say of me, "Because his throat is dry he ought now to take a drink," you would be expressing a plan of something like the following form: *if in his shoes, take a drink*.

It is important to see that the kind of plan Gibbard has in mind here is different from an ordinary contingency plan.[7] Normally, contingency plans

involve imagining oneself in a hypothetical situation and settling on what to do in that situation. For Gibbard's purposes, though, the situation in question needs to be understood extremely broadly. Even the agent's identity is part of the situation. Otherwise, difference in plans is no guarantee of disagreement. To see this, suppose, in the example of taking a drink, that Jeremy plans *if in his shoes, take a drink* and Alice plans *if in his shoes, don't take a drink*. It is important for Gibbard that these two plans be incompatible, which is to say that Jeremy and Alice actually disagree about what to do. If Jeremy can draw a distinction between being himself in my shoes and being Alice in my shoes, he can then conclude both *if (Jeremy) in his shoes, take a drink*, and *if Alice in his shoes, don't take a drink*. The same goes for Alice, in which case their plans could be perfectly compatible with one another. They could, in fact, agree on all of their planning judgments. To avoid this, Gibbard uses a very broad sense of what it is to be in someone else's shoes. To be in my shoes is not just to find oneself in a situation like mine, but actually to be me in the situation I am in. That way, when Jeremy plans *if in his shoes, take a drink* and Alice plans *if in his shoes, don't take a drink* their plans are incompatible, which is to say that they have a genuine disagreement in plans. Gibbard calls plans of this sort "hypothetical plans" as a way of distinguishing them from more ordinary "contingency plans" that involve settling on what to do as oneself in non-actual (but possible) circumstances.[8]

Given this understanding of hypothetical planning, Gibbard's general approach is to begin with a concept that we are forced to understand expressivistically, the concept of the thing to do, and then argue that our more familiar normative concepts are best understood in terms of this expressivistic concept, which is to say that they are expressions of hypothetical plans. The expressivist analysis is therefore supposed to spread throughout all normative discourse.

This is an appealing and powerful strategy for the defense of expressivism, but I will argue that it is not fully successful, and I will further suggest that the way in which it falls short of its goal is indicative of a problem faced by expressivism more generally. The structure of my argument, as indicated above, will be to demonstrate the limitations of Gibbard's view in the context of discussing two forms of innocent mistakes.

II. PURE PLANNING INNOCENT MISTAKES

One of Gibbard's strategies for capturing the richness and complexity of our actual normative judgments in terms of hypothetical plans is by drawing a distinction between what one plans to do and what one plans to plan to do. This is intended to make room for partial endorsement of a course of action, along the lines of the more familiar partial endorsement involved in the judgment that someone has made an innocent mistake that arises out of an epistemically

unprivileged situation. Such mistakes, which I call epistemic innocent mistakes, are common, and their existence should not be controversial. Any time I make a decision or form a plan on the basis of a false, warranted belief, I am liable to make an innocent mistake of this sort. In addition to these everyday epistemic innocent mistakes, however, Gibbard believes it is possible to make an innocent mistake in planning that has no ground in an innocent mistake about facts.

The sort of mistake Gibbard has in mind becomes possible when there is a divergence between the thing to do and the thing to plan to do. If you are in a situation where these two come apart, I can coherently (and correctly) endorse your plan and condemn your action.[9] I can say of you both that your plan was well formed (that the thing you planned to do and did was the thing to plan to do), and that in following through on your plan you erred (that the thing you planned to do and did was not the thing to do).

It is important to recognize that in determining whether such pure planning innocent mistakes are possible, we are most immediately engaged with a normative rather than metaethical question. We need to decide whether there are situations in which the thing to plan to do and the thing to do come apart, and that depends on the plans we adopt. Gibbard treats this issue as a second-order normative question, which is to say that he defends the possibility of pure planning innocent mistakes by offering an appealing strategy for forming plans

under which one's judgments of the thing to do and the thing to plan to do will occasionally diverge.

My response to Gibbard's analysis will come in two stages. First I will engage with his second-order normative account and attempt to show that there is a more plausible strategy for planning than the one he advocates. Under my preferred strategy for planning, plans for what to do and what to plan to do will always align, and we will therefore never attribute pure planning innocent mistakes. Continuing this second-order normative debate, I will also address some of Gibbard's earlier work in which he offers a more fully developed example purporting to show that an ordinary, acceptable normative stance can lead to the attribution of pure planning innocent mistakes. In the second stage of my response, at the end of this section, I will step back from these second-order normative questions and argue that the role of planning judgments in Gibbard's expressivist metaethics speaks against the possibility of pure planning innocent mistakes. There I will be offering a metaethical argument, but one that has bearing on the (seemingly) normative question of what plans to form. In the following section I will then demonstrate a genuine form of partial endorsement that Gibbard might have hoped to capture through attributions of pure planning innocent mistakes. My ultimate conclusion will be that Gibbard's view is unable to account for the form of innocent mistake related to this genuine partial endorsement.

To begin with, though, let me focus on Gibbard's presentation of a seemingly plausible strategy for planning which he thinks leads to attributions of pure planning innocent mistakes. His idea is this. In planning, look for people to trust, people to whom we can defer with confidence. Then plan like them.

To see how this policy leads to the attribution of pure planning innocent mistakes, let us follow Gibbard's lead and consider the situation of being Socrates in the prison cell choosing between fleeing Athens or drinking hemlock.[10] Call this situation S. In thinking about what to do in S I will take Socrates and Xanthippe as hypothetical planners (or hypothetical agents). Let me call them Soc and Xanti for short. Now suppose that Soc is such that if he were in ideal conditions he would judge that the thing to do is drink the hemlock, but Xanti is such that if she were in ideal conditions she would judge that the thing to do is flee. Let me designate the idealized Soc with "Soc+" and the idealized Xanti with "Xanti+." I am going to be vague about what exactly it means for an agent to be idealized. One possibility is that to be idealized means to confront all relevant considerations vividly and repeatedly in a clear, dispassionate manner. That characterization may need modification, but the central point is that idealization leads to better judgment, so that it makes sense for an agent to trust an idealized version of himself.[11]

Recall that Gibbard's suggestion for how to plan is to identify people who can be trusted, people to whom we are willing to defer, and then plan like them.

Suppose that Xanti adopts this policy in considering the situation S of being Soc and choosing between drinking hemlock or fleeing. She says to herself, *What shall I plan to do if in S? Well, the best way to plan is by looking for people to trust. If in S, it makes sense to trust Soc+. So, if in S, the thing to plan to do is whatever Soc+ plans to do. Soc+ plans to drink the hemlock. In that case, if in S, let me plan to drink the hemlock.* Then she asks herself directly what is the thing to do if in S. Now she thinks, *What shall I do if in S? Well, to settle on what to do is to form a plan, and the best way to do that is by looking for people to trust. It makes sense for me to trust Xanti+. So the thing to do in S is whatever Xanti+ plans to do in S. Xanti+ plans to flee. So, if in S, let me flee.* Xanti, then, plans to flee if in S, yet also plans to plan to drink the hemlock if in S. She is now in position to say that although Soc planned correctly in planning to drink the hemlock, drinking the hemlock was not the thing to do. If she believes both of these things, she holds that Soc made a pure planning innocent mistake, because to hold that he made a pure planning innocent mistake just is to endorse his plan to drink the hemlock but disagree with his action of drinking the hemlock.[12]

Although I will ultimately argue that there is a deep tension between the role of hypothetical plans in Gibbard's analysis and his claim that it makes sense for Xanti to attribute a pure planning innocent mistake to Soc, I do concede that it is at least possible for Xanti to follow Gibbard's proposed strategy and end up with a divergence between what she plans to do in S and what she plans to plan to

do in S. It is therefore important that I identify an alternative strategy for planning which does not lead Xanti to the attribution of a pure planning innocent mistake. One obvious possibility would be for Xanti to defer consistently to an idealization of her actual self. She already does this, on Gibbard's view, in considering the thing to do in S. Why not do it as well in considering the thing to plan to do in S? This way she would determine the thing to do in S and the thing to plan to do in S the same way.

There is, however, a legitimate question as to whether this amounts to a coherent strategy for planning. Essentially, what Xanti would be saying would be that in determining how to plan in S, she should consult her idealized actual self and plan accordingly. But who is an idealized Xanti to Soc? Remember, we are talking here about what Gibbard calls hypothetical planning. Xanti is therefore supposed to be making a plan for how to plan in the case of being Soc in S. She is not making a more familiar contingency plan for how to plan if herself in a situation similar to the one Soc faces in S. As a result, if her plan is 'defer to my actual self idealized', there is a real question of how, as Soc in S, she is supposed to identify the guru in question. Without some refinement, this is akin to planning "Buy low, sell high." In other words, it suffers the defect of being a plan that is not formulated in recognitionally available terms. The problem with "buy low, sell high" is that one doesn't know if the current price is low or high until after the fact. Similarly the problem with "if Soc in S, defer to an idealized version of my

actual self," which in this case is Xanti, is that Soc in S will have no way to identify Xanti as the "actual self" of that plan.[13]

As a second attempt, we can circumvent this problem by having Xanti think "if Soc in S, defer to Xanti+" rather than "if Soc in S, defer to an idealized version of my actual self." This avoids the problem of Soc being unable to identify Xanti+ as the referent of the phrase "idealized version of my actual self," but it raises another sense of the question: Who is an idealized Xanti to Soc? Here the worry is, if Soc in S, why care what Xanti+ thinks? Why take her as someone to be trusted, someone to whom it makes sense to defer?

I take this to be a genuine concern, and I will return to it in Section III, but for now let me point out that it is at least coherent for Xanti to plan to defer to Xanti+ in S, and that if she does so in determining both what to do and what to plan to do she will not think Soc is in a position to make a pure planning innocent mistake. Moreover, there is another way that Xanti could avoid attributing a pure planning innocent mistake to Soc in S. Rather than deferring to Xanti+, she could instead settle on deferring to Soc+ to determine both what to do and what to plan to do in S. In that case, she would conclude that the thing to do is drink the hemlock (because Soc+ would advise her to drink) and the thing to plan to do is plan to drink the hemlock (again, because of the advice of Soc+). Either way, whether Xanti chooses to defer to Xanti+ or Soc+, her conclusion about what to

do will match her conclusion about what to plan to do. She will therefore not attribute a pure planning innocent mistake to Soc.

Moreover, these methods of planning are actually just instances of a more general set of strategies under which Xanti will deny the possibility of pure planning innocent mistakes. Whether to adopt such a strategy is, at least at first glance, a question of second-order normative ethics. Viewed in this way, the general argument for adopting such a strategy goes as follows. Whenever one considers a specific case, such as being Soc in S, one must confront the question of whether, and to what degree, one is to defer to the judgments of Soc+ in determining what to do in S, as opposed to deferring to the judgments of an idealized version of oneself. Say one decides to defer to Soc+ to extent ε.[14] Now suppose one is further wondering what to plan in S. One is faced with a similar question. Should one defer to Soc+ in determining the thing to plan to do in S? My proposal is that one should defer to Soc+ to extent ε here as well. The central idea is that whatever reasons there might be to defer to Soc+ in deciding what to do if Soc in S would also tell one to defer to Soc+ in deciding what to plan to do if Soc in S. If this is right, pure planning innocent mistakes will not arise, because there will be no way to pry apart judgments of the thing to do and judgments of the thing to plan to do.

As I see it, Gibbard's strongest argument in favor of permitting pure planning innocent mistakes, comes in *Wise Choices, Apt Feelings*, where he

purports to offer a case in which a relatively simple normative stance leads to the attribution of pure planning innocent mistakes.[15] The argument in *Wise Choices* is framed not in terms of planning judgments, but rather in terms of two types of higher order norms that tell us what norms to accept, norms of rationale and norms of warrant. *Norms of rationale* tell us to accept norms because they play a certain role. The example Gibbard gives is, "Accept a norm if its acceptance, in one's community, would most enhance a sense of meaning in life."[16] *Norms of warrant*, on the other hand, tell us what process to use in determining which norms to accept. Here Gibbard's example is a dialectical equilibrium theory that says to accept whatever norms persist in dialectic equilibrium.

Gibbard argues that it is perfectly consistent to hold norms of rationale and norms of warrant that seem to conflict. He imagines a community of Greeks who hold perfectionist norms of rationale and whose norms of warrant endorse dialectical equilibrium. He further imagines that when another group, the Scythians, reaches dialectical equilibrium they adopt hedonistic norms of rationale. The catch is that perfectionism dictates that the Scythians be warlike and hedonism dictates that they be peaceful. Gibbard writes, "Greek norms, then, tell a Scythian to accept norms that prescribe peace, but to choose war."[17] Translating this into our planning language, the Greek norms say roughly this:

(A) *What to do if a Scythian? Be a soldier.*

(B) *What to plan if a Scythian? Plan to be a pacifist.*

(A) stems from the Greeks' perfectionist norm of rationale, together with the fact that perfectionism dictates that the Scythians be warlike. (B) stems from the Greeks' norm of warrant, together with the facts that the Scythians endorse hedonism in dialectical equilibrium, and that hedonism dictates that they be pacifists.

It is the combination of (A) and (B) that leads the Greeks to attribute a pure planning innocent mistake to the Scythians. What I want to show is that on any understanding of the Greeks that genuinely commits them to (A), they ought to reject (B), and vice versa.

In order to defend the combination of (A) and (B), Gibbard examines the interplay between the Greek norms of rationale and warrant. He identifies two possible relationships between them. The first, which he finds unhelpful, is to suppose that the Greeks ground their norms of rationale on their norms of warrant. The problem here is that this would lead the Greeks to acknowledge that people who reach different conclusions in dialectic equilibrium could have alternate norms of rationale. The perfectionist rationale would be relative to their community, and it would not dictate that the thing to do if a Scythian is to perfect oneself by becoming a soldier. On this approach, then, the Greeks would be led to reject (A).

The other possibility, which Gibbard thinks will vindicate the Greek commitment to both (A) and (B), is to suppose that the Greeks ground their norms of warrant in their norms of rationale. Gibbard writes:

> The Greeks indeed can have a rationale for thinking for themselves and resisting outside influence. Thinking together in one's narrow community, they might say, is the best way to develop one's human capacities. That is what recommends it, and this perfectionist rationale, Greeks can admit, gives Scythians, too, good reason to think for themselves. Alas, though, if Scythians think for themselves they come to reject perfectionism. …They have applied the right norms of warrant, norms supported by a good rationale. In so doing, they have come to reject that very rationale. In this they are mistaken.[18]

The idea here is that we begin by taking the perfectionist norm of rationale as basic. This underwrites (A). In addition, the perfectionist rationale supports endorsing the results of dialectical equilibrium as warranted. Moreover, in dialectical equilibrium the Scythians reject perfectionism for hedonism, and conclude that being peaceful is the surest route to happiness. This leads to (B). Gibbard's conclusion is that the Greeks can consistently, and even appropriately,

assert both (A) and (B), as long as we suppose that they ground their norms of warrant on perfectionist norms of rationale.

If this argument works, we have an example of a simple, straightforward normative view, perfectionism, leading to the attribution of pure planning innocent mistakes. That would make my claim that we should only adopt a normative view that denies the possibility of such mistakes highly implausible.

There is, however, reason to believe Gibbard's analysis does not vindicate the Greek view as intended. Recall that on Gibbard's solution the Greeks endorse the result of dialectical equilibrium not because of some fundamental norm of warrant but because doing so is the best way to develop one's human capacities. That is, their norms of warrant are grounded in a more basic perfectionist rationale. Earlier, however, the Greeks are said to believe that the best way for the Scythians to develop their human capacities is for them to be warlike. If the Greeks stick to the initial understanding of Scythian perfection, the Greeks will conclude that the Scythians have no rationale for thinking for themselves. Instead, the perfectionist rationale supports different norms of warrant for the two communities. The thought, *What to plan if a Scythian? Plan to be a pacifist*, would have no basis. That is, the Greeks would be forced to reject (B).

Of course, the Greeks could also reconsider their view that perfectionism dictates that Scythians be warlike. They might instead maintain that thinking for oneself is the best way for any people to develop their human capacities,

regardless of the result. This would allow them to retain (B) *What to plan if a Scythian? Plan to be a pacifist*, but it would undermine (A) *What to do if a Scythian? Be a soldier*. So it seems that taking norms of warrant as grounded in norms of rationale will not justify the Greeks in maintaining both of their attitudes towards the Scythians at once. The Greeks, as perfectionists, are therefore not committed to attributions of pure planning innocent mistakes, and the second-order normative position that we ought not to hold normative views that lead us to attribute such mistakes remains a viable option.

There is one other possible interpretation of Gibbard's discussion of the Greeks and the Scythians that is worth mentioning. Perhaps the idea is that, according to the Greeks, it would be most perfect if the Scythians were to engage in dialectical equilibrium and thereby endorse perfectionism and go on to perfect their warlike qualities. This seems odd, though, given that it has already been stipulated that when the Scythians engage in dialectical equilibrium they endorse hedonism. If we give up that stipulation, the Greeks are free to think (B') *What to plan if a Scythian? Plan to be a soldier (through dialectical equilibrium).* This could be combined with (A) *What to do if a Scythian? Be a soldier*, without leading to the attribution of a pure planning innocent mistake. If, however, we retain the stipulation that the Scythians endorse hedonism when in dialectical equilibrium and rule out (B'), the Greeks must refine their notion of what perfection is for a Scythian and either give up (A) because perfectionism dictates

thinking for oneself regardless of the results or give up (B) because perfectionism only dictates thinking for oneself if such thought will lead one to perfect oneself. Either way, they will not be committed to attributing pure planning innocent mistakes to the Scythians.

If I am right that the argument from *Wise Choices* does not work, we are back to the thought that whether to attribute pure planning innocent mistakes is a live question. As indicated earlier, I think there is a powerful second-order normative argument against attributing pure planning innocent mistakes, one that is grounded in the thought that whatever makes an action the thing to do should also make it the thing to plan to do, and vice versa. Moreover, I will now argue that attributions of pure planning innocent mistakes are inconsistent with the role that hypothetical plans are supposed to play in Gibbard's view.

This inconsistency stems from the fact that determining a certain action to be the thing to do in a given situation involves identifying it as the choiceworthy action in that situation. Unless we give up the claim that this action is choiceworthy, that it is the thing to do, it makes no sense to do something else.[19] But attributing a pure planning innocent mistake just is both (i) planning to judge an action choiceworthy (planning to plan to do it) and (ii) planning to do something else, without in any way repudiating the judgment of choiceworthiness. The point is not that we must think we will always be correct in our identification of the choiceworthy action in a situation, just that it makes no sense to plan to do

something else without also revising our view of the choiceworthiness of the initial action. If plans don't work this way, then I begin to lose my grip on what the point of planning is in the first place. That is, if I think I can consistently plan to do one thing and then do something else, without in any way revising or repudiating the plan, then I no longer understand what plans are, or why I should bother with them at all. If plans are to function as decisions about what to do in hypothetical situations, which is the role Gibbard assigns to them, we therefore should not plan in a way that permits pure planning innocent mistakes.

It is worth taking note of a potential objection here that could arise out of the literature on whether there can be reasons to form an intention to do something that do not also serve as reasons to perform the act in question.[20] Consider Gregory Kavka's toxin puzzle, in which an agent is offered a prize for forming an intention to drink a toxin that will cause one day of painful illness.[21] The prize is large enough that the agent is willing to suffer the effects of the toxin in order to win it, but the catch is that the prize is tied not to drinking the toxin but to intending to drink the toxin, and is awarded before the toxin is actually consumed. In this situation, the existence of the prize seems to provide a clear reason to intend to drink the toxin, but no reason to follow through and actually drink it, given that the prize will have already been awarded (or not).[22] Presumably a similar sort of case could be developed for Gibbard's plans, which suggests that reasons to plan to do something can diverge from reasons to do it. If

this is right, then perhaps we should follow Gibbard and endorse pure planning innocent mistakes.

Notice, though, that the key to cases of this sort is that there are consequences of planning (or intending) that are independent of the consequences of performing the planned (or intended) action. This is what drives the potential wedge between the thing to plan to do and the thing to do. In the cases I have been discussing, however, there are no such consequences in play. Gibbard does not recommend that Xanti charge Soc with having made a pure planning innocent mistake on the grounds that Soc's planning to drink the hemlock provides benefits that can be reaped without actually drinking the hemlock. Rather, Gibbard's view is that the attribution of pure planning innocent mistakes results from an acceptable, or perhaps even optimal, strategy for planning. He thinks Xanti should defer to different authorities when deciding what to plan to do and what to do, and this does not depend on there being a gap between the consequences of forming a plan and those of performing the planned action.

Moreover, even if there were effects of planning that were independent of the effects of performing the planned action, there is a further difficulty with defending pure planning innocent mistakes through the use of cases such as Kavka's toxin puzzle. After all, there is widespread agreement in discussion of the toxin puzzle that a rational agent, provided she has no recourse to external incentives and no way to circumvent her rationality, will find it psychologically

impossible to intend to drink the toxin while at the same time recognizing that, even after having formed the intention, the balance of reasons will speak against actually drinking it.[23] With this as a starting point, the debate over the toxin puzzle centers on whether Kavka was correct to deny that the agent can create a reason to drink the toxin through the act of decision or intention formation. The shared idea is that part of what it is to intend to drink the toxin is to believe that one will drink it, or at least attempt to drink it.[24] By analogy, part of what it is to form a plan is to believe that one will follow through on it, or at least attempt to follow through on it. This again leads to the idea that it makes no sense to plan to do one thing and then, even while maintaining that plan, do something else. Drawing an analogy between intentions and Gibbard's plans therefore reinforces, rather than undermines, the idea that we should deny the possibility of pure planning innocent mistakes.

Interestingly, the analysis Gibbard himself offers in a slightly different context further supports the rejection of pure planning innocent mistakes. When considering the interaction between judgments and plans for judgments, Gibbard writes:

> If I judge that cattle are on the hill, then I'm committed to the plan to judge, if in my present situation, that cattle are on the hill. It would be incoherent, after all, to make a judgment and yet, in my

> plans, rule out so judging in that very situation. …Not, to be sure, that I normally go out of my way to make such a plan; a requirement that I must would lead to a regress: must I plan so to judge, plan to plan so to judge, and so on? Still, I do *commit* myself to all these layers of plans: I rule out rejecting them.[25]

This is precisely the line of thought, applied to the relationship between what I plan to do and what I plan to plan to do (rather than the relationship between what I plan to judge and what I plan to plan to judge), that leads me to disallow pure planning innocent mistakes. Insofar as Xanti plans to flee if Soc in S, she has committed herself to planning to plan to flee if Soc in S, and has ruled out planning to plan to drink the hemlock if Soc in S. If she violates this commitment, her plans become incoherent. In the passage above, Gibbard rejects this sort of incoherence with respect to plans regarding judgments. I think he is right to do so, because embracing such incoherence would make it impossible for plans to play the role of decisions regarding hypothetical scenarios. Similarly, if we want to maintain Gibbard's general picture of how plans function, we should reject the possibility of pure planning innocent mistakes regarding what to do.

III. INNOCENT MISTAKES OF MENTAL CONSTITUTION

In the earlier discussion of Xanti's plans for what to do and what to plan to do in S, I suggested a range of possible strategies she might adopt that would allow her to avoid attributing a pure planning innocent mistake to Soc. The idea was that she should defer to Soc+, as opposed to Xanti+, to the same degree in settling what to do if Soc in S as in settling what to plan to do if Soc in S. I called her degree of deference to Soc+ "ε," which was intended to range over possible values from 0, which would indicate total deference to Xanti+ and total disregard for Soc+, to 1, which would indicate the opposite: total deference to Soc+ and total disregard for Xanti+.

Let me now consider an argument in favor of a large ε, perhaps even an ε of 1. The pressure towards a large ε stems from the worry, mentioned earlier, that it may not make very much sense for Soc in S to choose Xanti+ over Soc+ as his guru. I do not mean to suggest that we must always take ourselves, or even idealized versions of ourselves, as the ultimate authority in questions of what to do. On the contrary, it is quite reasonable to identify another as what we might call an "evaluative expert," and take her advice as weighing heavily in our deliberation, or perhaps settling it altogether. In order for this to be reasonable, however, there must be some identifiable feature of this advisor in virtue of which we choose to defer to her judgment. We must at least have the authority to choose for ourselves between competing advisors. No advisor can help us with that choice.

It would, therefore, be perfectly reasonable for Soc to choose Xanti+ over Soc+ as an advisor if he could identify some characteristic of Xanti+ that would make her advice especially valuable. It looks, however, as though the reason for Xanti to want to defer to Xanti+ if Soc in S is not that Xanti+ has some such characteristic. Rather, Xanti is inclined to defer to Xanti+ simply because Xanti+ is an idealization of her actual self, and in that case the deference to Xanti+ makes no sense from the perspective of Soc in S, which is the perspective Xanti must adopt in forming a hypothetical plan regarding what to do (or plan to do) in S. We could of course tell the story differently. Perhaps Xanti+ has some genuine, identifiable advantage over Soc+ that makes her advice particularly trustworthy. It is important to keep in mind, though, that this would have to be an advantage recognizable by Soc in S. The general conclusion is that in planning for a situation it makes sense to defer to an idealized version of the agent in that situation, unless the agent in the situation can himself identify some other advisor as more trustworthy.

This, however, leads to some peculiar results. Consider, for instance, the situation M of being a psychopathic murderer, call him Mortis, who would continue to prefer to murder even if idealized. That is, both Mortis and Mortis+ prefer to murder, and let us further suppose that Mortis has no way of identifying a more trustworthy advisor than Mortis+.

The question now is: what to say about situation M? Notice that the argument just given commits me to what may look like the implausible conclusion that the thing to do, and the thing to plan to do, in M is murder. I think this actually is the right conclusion, as far as judgments of the thing to do and the thing to plan to do go, but what it shows is that there is a critical gap between such judgments and the common sense normative judgments that Gibbard wants them to capture. After all, what should one say about M? It seems to me that, given the extreme nature of the case, we are forced to concede that it makes sense for Mortis to murder in M. It would be unreasonable to deny this, given that Mortis's mental constitution is such that he has no possible access to a perspective from which he could recognize what is wrong with murdering. In other words, in spite of whatever else we might want to say about M, we must admit that, for Mortis in M, murdering is the thing to do. The same argument applies when thinking about the thing to plan to do in M, and again the conclusion is that, for Mortis in M, murdering is in fact the thing to plan to do.

I also want to suggest, however, that this is not *all* that we can say about situation M. Rather, I think it makes sense to make a further claim, which is that even though murdering is the thing to do for Mortis in M, it is still *wrong*. That is, the case of Mortis's murdering in M is an instance of what I will call an innocent mistake of mental constitution. M is a situation in which the thing to do and the thing to plan to do come apart from our full normative assessment of

Mortis's action, or so I claim. In other words, there is a sense in which Mortis's action is wrong, even though it is both the thing to do and the thing to plan to do. This is possible because Mortis is constituted in such a way that he is incapable of recognizing the wrong-ness of his action. In murdering, Mortis therefore commits an innocent mistake of mental constitution.

One might wonder, at this point, whether this sort of case is even possible. That is, could there be a person like Mortis, who truly would prefer to murder even when deliberating under ideal conditions? If we could rule out such a possibility, then the problem I am posing might dissolve. Nonetheless, I think Gibbard would be reluctant to explain away the problem in this manner. He makes room for cases such as Mortis in his discussion of what he calls "constitutional impasse," in which he acknowledges that there can be disagreement between idealized planners.[26] In fact, it is precisely the possibility of such disagreement that leads Gibbard to reject a straightforward ideal observer view.[27] The disagreement between M+ and, presumably, the idealized version of most of us is certainly more significant than the disagreement between Xanti+ and Soc+ discussed earlier, but I take it that Gibbard is committed to the theoretical possibility of both. Moreover, even if we were to reject Mortis as too far fetched to take seriously, my objection could be formulated in terms of the disagreement between Xanti+ and Soc+, with Xanti attributing to Soc an innocent mistake of mental constitution when she endorses his drinking of the hemlock and his plan to

drink the hemlock but nonetheless believes that drinking the hemlock is in some sense wrong. I find thinking in terms of Mortis helpful because it emphasizes the gap between our judgments of what to do and plan to do in M and our overall assessment of Mortis' actions, but readers who balk at the example can focus instead on Xanti's attitudes towards Soc in S without impeding the argument.

As I see it, what we should say about Mortis in M, as well as what Xanti should say about Soc in S, demonstrates a specific sense in which Gibbard's expressivism lacks the resources to capture an important aspect of moral objectivity. On Gibbard's view, to conclude that murdering is both the thing to do and the thing to plan to do in M is to come to a full-scale endorsement of Mortis's murdering. Or more precisely, Gibbard's view commits us to move from the thought that it makes sense for Mortis to murder and plan to murder, and that it would continue to make sense for Mortis to murder and plan to murder even if Mortis were in ideal epistemic conditions and ideally placed to judge whether to murder, and that Mortis has no way to identify a more trustworthy guru than Mortis+, to a full-scale endorsement of Mortis's murdering. Once we have reached the conclusion that murdering is both the thing to do and the thing to plan to do in M, Gibbard's analysis of normative judgments in terms of planning judgments leaves no room for us to maintain that Mortis's murdering in M is wrong. On his view, innocent mistakes of mental constitution are incoherent. Gibbard's analysis therefore commits us to what I would count as an unacceptable

moral relativism by denying us the possibility of criticizing Mortis's act of murder.

It is important to recognize that this problem with the version of expressivism found in *Thinking How to Live* is symptomatic of a more general difficulty for expressivist projects. Before exploring that connection, however, let me consider some possible responses on Gibbard's behalf. Notice that I have argued, in Section II, that Gibbard's view incorrectly permits pure planning innocent mistakes, which I claim to be illusory. Now I am arguing that there is a different form of innocent mistake, an innocent mistake of mental constitution, and that although these mistakes are genuine parts of the normative landscape, Gibbard's view cannot capture them. It might therefore appear as though Gibbard's response should be to point out that if we reject both of my arguments the problem will go away. The suggestion here would be that by allowing for pure planning innocent mistakes Gibbard may be able to capture the partial endorsement that I am associating with a recognition of innocent mistakes of mental constitution.

Moreover, Gibbard has a further defense of pure planning innocent mistakes that I have not yet addressed that looks particularly relevant here. Gibbard's claim is that in situations like M, a strategy of deferring to Mortis+ in determining both what to do and what to plan to do will be unacceptably alienating.[28] Given such deference, one's own views about murder matter only to

the extent that Mortis is capable of recognizing and agreeing with them. Gibbard suggests finessing this alienation by combining deference to an idealized version of oneself in determining what to do with deference to an idealized version of Mortis in determining what to plan to do, much as Gibbard would have Xanti plan for the case of being Soc in S. In other words, Gibbard's suggestion is that we can avoid alienation and make room for a partial endorsement of Mortis' murdering by attributing to Mortis a pure planning innocent mistake.

My response to this worry about alienation is that it actually makes sense to be alienated from oneself when planning *for the case of being Mortis*. More typical contingency plans take the form of planning what to do *as oneself* in some hypothetical situation, and there alienation would be worrisome, but hypothetical planning for the case of being Mortis just is planning for the case of being an alien with respect to oneself, so alienation is not necessarily a problem. Moreover, insofar as alienation is a genuine concern even when deliberating about situations in which one is not oneself, there seems to be no reason to find such alienation more troubling when thinking about what to do than when thinking about what to plan to do. At most, the worry about alienation should lead to embracing a smaller ε, which is to say deferring less to Mortis+ both in thinking about what to do and in thinking about what to plan to do.

Furthermore, even if we were to think it made sense to permit pure planning innocent mistakes, perhaps in order to prevent alienation, an appeal to

such mistakes would not provide an adequate response to my objection concerning objectivity. I concede that if we were to allow pure planning innocent mistakes it would indeed be possible to criticize Mortis from within Gibbard's expressivistic view. Notice, though, what the criticism would be. Following Gibbard's analysis of how Xanti thinks about the case of Soc in S, we might conclude that the thing for Mortis to plan to do is plan to murder, but that the thing for Mortis to do is not to murder. This result would come from deferring to Mortis+ in determining what to plan, but deferring to an idealized version of oneself in determining what to do. I grant that this would allow us to hold back from a full-scale endorsement of Mortis's murdering, but not, it seems, in the right way. The challenge of Mortis's situation is that it actually does make sense for Mortis to murder, just as it makes sense for Mortis to plan to murder. What we really want to say, and what Gibbard's view makes us incapable of saying, is that in spite of these conclusions about the thing to do and the thing to plan to do if Mortis in M, Mortis's murdering is still wrong.

To see perhaps more clearly that Gibbard's analysis deprives us of the ability to make this claim, it may help to draw a comparison between innocent mistakes of mental constitution and more familiar epistemic innocent mistakes. If we are considering a situation, call it E, in which the agent is epistemically limited so that the best available evidence would lead her to do the wrong thing, we would presumably conclude that the thing to do in E, because of the limited

information, is to err. In Gibbard's terms, we would plan if in E to do some action A, even though we think doing A would in some sense be a mistake. In what sense would it be a mistake? Presumably it would be a mistake because if we were planning for a similar situation but with full information we would plan not to do A. This may be somewhat of a simplification of how to understand ordinary epistemic innocent mistakes, but some such account must hold in order for Gibbard to be able to interpret as planning judgments both our endorsement of doing A in E and our sense that doing A in E is a mistake.

Compare this with an innocent mistake of mental constitution. If we are to conclude that Mortis makes such a mistake in M (or if Xanti is to conclude that Soc makes such a mistake in S), then we are concluding that in that sort of situation, because of the sort of person we would be in the situation, we should err. Again, though, we must ask ourselves in what sense the action in question counts as an error, or a mistake. I do not mean to imply that Mortis's murdering is not a mistake, or that its being a mistake is incompatible with Mortis having most reason to murder. Rather, my point is that it *is* a mistake in some sense, even though murdering is the thing to do and the thing to plan to do in M, and that Gibbard's planning language is unable to capture this normative judgment. Analogously, Xanti thinks that it is a mistake to drink the hemlock, even though she concludes that drinking the hemlock is the thing to do and the thing to plan to

do, and my objection is that Gibbard cannot capture the sense in which she withholds full endorsement of drinking of the hemlock.

It might seem as though this objection is easily met, in that Gibbard could simply claim that when I say that Mortis' murdering in M is wrong I am expressing my plan not to murder if I were in a situation similar to Mortis. It wouldn't be exactly the situation of Mortis, because in that situation I would plan to murder, given that I would have the mental constitution of Mortis and would defer to Mortis+, who would endorse murdering. But I can also imagine a situation just like that of Mortis except that the agent has *my* mental constitution, and I could quite reasonably plan not to murder in that situation, because an idealized version of *me* would not endorse murdering.

This solution, however, comes at a substantial cost. After all, if my judgment that Mortis' murdering is wrong depends on my plans for what to do if in Mortis' situation but with my mental constitution, then your judgment that Mortis' murdering is wrong depends on your plans for what to do if in Mortis' situation but with your mental constitution. In that case, even if we both assert that Mortis does something wrong in murdering, we are not really agreeing with one another, because we are expressing plans for two different situations. More problematically, if someone who shared Mortis' mental constitution were to endorse Mortis' murdering in M, he would not be disagreeing with either of us, because he would be expressing a plan for yet a third situation.[29] The trouble is

that if our overall assessment of Mortis' behavior is indexed to our own mental constitutions, then an apparent difference in our overall assessment does not involve an actual disagreement in plans, which on Gibbard's view implies that it does not involve a normative disagreement at all. The need to avoid this sort of problem, and ensure that difference in plans amounts to genuine disagreement, is precisely why Gibbard's analysis is focused on hypothetical plans for the case of being someone else rather than contingency plans for being oneself in a situation roughly like theirs.

Perhaps the most obvious way for Gibbard to attempt to capture our judgment that Mortis' murdering is wrong, or Xanti's judgment that Soc's drinking the hemlock is wrong, would be through an appeal to his norm-expressivist analysis of morality presented in *Wise Choices, Apt Feelings*.[30] The idea would be that we might prescribe guilt on Mortis' behalf, and resentment on the part of others, and that these norms governing appropriate guilt and resentment are what capture the sense in which we think Mortis is wrong to murder, even though we concede that murdering is the thing to do in M. In the planning language of *Thinking How To Live*, we would plan to murder in M, and also to feel guilt about murdering in M, and, as ourselves, to resent Mortis's murdering in M.

Let me focus first on resentment, because I think it is easier to see why norms governing resentment are incapable of distinguishing between partial and

full endorsement of M's murdering. The problem is that even if I believe that Mortis' murdering is wrong, I can nonetheless conclude that whether to resent Mortis depends on one's mental constitution. Suppose I imagine Portis, who shares Mortis' mental constitution, and I consider whether, as Portis, to resent Mortis. The same line of thought that led me to conclude that murdering is the thing to do in M, given Mortis' mental constitution, also generates the conclusion that, as Portis, resentment is not the emotion to feel towards Mortis. That is to say, in planning what to feel for the case of being Portis I defer to Portis+, and because Portis+ sees nothing objectionable in Mortis' behavior, Portis+ does not recommend resentment. Moreover, and this is the critical point, it is appropriate for me to reach this conclusion even though I continue to maintain that Mortis' murdering is wrong. As a result, my plans for resentment cannot be used to capture the sense in which I withhold full endorsement of Mortis' murdering. Similarly, even if Xanti thinks Soc is wrong to drink the hemlock and resents him for it, as long as she also concludes that drinking the hemlock is the thing for Soc to do, given who he is and what he is like, she has no basis for planning to resent Soc if she were someone else, say Plato, who shares Soc's mental constitution.

This failure of plans for resentment to capture the sense in which we withhold endorsement of Mortis' murdering (or Xanti withholds endorsement of Soc's drinking the hemlock) stems from the fact that plans for resentment, like plans for behavior, are limited by differences in mental constitution. As a result,

plans for resentment will not provide for genuine normative disagreement in cases involving differences in mental constitution. Someone who shares Mortis' mental constitution might plan to resent Mortis if like the rest of us, yet plan not to resent Mortis himself. He would, therefore, endorse the same norm for resentment that we do. Similarly, Plato and Xanti might agree both that it makes sense for Xanti, given her mental constitution, to resent Soc for drinking the hemlock and that it makes sense for Plato, given his mental constitution, not to resent Soc. In that case, their norms of resentment are identical. Nonetheless, there is a sense in which Xanti wants to condemn Soc's drinking of the hemlock and Plato wants to endorse it, and there is a sense in which they genuinely disagree about this. It is this disagreement, in which each is left wanting to say something like "but it really is wrong" or "but it really is not wrong," that Gibbard is unable to capture either through plans for what to do if Mortis or through plans for whether to resent Mortis.[31]

What about guilt? It may seem as though focusing on guilt provides Gibbard with the resources to capture the normative judgment in question. After all, unless we presuppose that moral reasons override non-moral reasons, we must acknowledge the possibility of cases in which there is most reason to do something that is nonetheless morally wrong. In these cases, it is perfectly reasonable to plan to do something and at the same time plan to feel guilt for doing it. Moreover, it looks as though M, on my analysis of the case, is just this

sort of situation. I have been insisting that it is appropriate to conclude that Mortis has most reason to murder in M, and yet that murdering in M is nonetheless wrong, and I have been challenging Gibbard to find a way to capture this set of judgments regarding M. It is natural to think Gibbard can meet this challenge by appealing to plans for guilt as follows. He might say that the difference between us (who condemn Mortis' murdering) and Portis (who doesn't) is that we plan to murder in M and to feel guilt over murdering in M, whereas Portis plans to murder in M and plans to feel no guilt. Similarly, if we suppose that Xanti takes suicide to be a moral wrong, Xanti would plan to drink the hemlock in S and feel guilty as she does so, whereas Plato would plan to drink it guilt-free.

This apparent solution, however, fails to take into account an important feature of the cases of Mortis and Soc. Perhaps the best way to see this is by drawing a distinction between three sets of possible judgments. Formulated generically, on might judge (a) X is the thing to do and it is not morally wrong. Alternatively, one might judge (b) X is the thing to do even though it is morally wrong, and I will be capable of recognizing it as morally wrong but still the thing to do. Or, as a third option, one might judge (c) X is the thing to do even though it is morally wrong, and I will be constituted so as to be incapable of recognizing it as morally wrong. As may now be evident, the apparent solution for Gibbard described in the previous paragraph is sufficient to distinguish (a) from (b), but

not from (c). Moreover, it is the distinction between (a) and (c) that is at stake in the cases I have been discussing.

Let me elaborate. When we translate (a) into Gibbard's planning language we get a plan to do X and to feel no guilt. When we translate (b) into Gibbard's planning language we get a plan to do X and to feel guilt. What about (c)? Here the same reasoning that has led to the conclusion that the thing for Mortis to do in M is murder, and that resentment is not the emotion to feel if Portis contemplating Mortis' murder, leads to the conclusion that guilt is not the emotion to feel for doing X. Sticking to the case of Mortis, the idea is that in planning whether to feel guilt in M, I defer to Mortis+ just as I defer to Mortis+ in planning whether to murder in M. Just as it is appropriate for me to conclude (perhaps begrudgingly) that the thing to in M is murder, it is appropriate for me to conclude that guilt is not the emotion to feel in M over murdering. And reaching this conclusion in no way requires me to give up my view that Mortis' murdering is still somehow wrong. I can therefore plan both to murder and not to feel guilt, and yet still withhold full endorsement from murdering. As before, it is this withheld endorsement, my judgment that Mortis' murdering really is wrong even though it is the thing to do in M and even though guilt is not the emotion to feel in M, that Gibbard's planning language cannot capture.

The problem for Gibbard is that (a) and (c), even though they are distinct normative assessments, translate into identical sets of plans regarding what to do

and what to feel. Once I adopt stance (c) towards Mortis' murdering, my plans become indistinguishable from those of Portis, who endorses Mortis' murdering wholeheartedly. Similarly, if Xanti adopts stance (c) towards Soc's drinking the hemlock, which is appropriate for her to do given the situation, she will plan to drink the hemlock guilt-free, just as Plato does, even though there is a genuine normative disagreement between them.

It is worth emphasizing that this problem for Gibbard is unrelated to the question of whether moral reasons are overriding. Views about the overridingness of moral reasons will come into play if we want to determine whether the normative stance contained in (b) is ever appropriate. Those who think moral reasons override non-moral reasons will deny that it ever makes sense to adopt stance (b) or the plans that go along with it. But that debate is independent of the objection I am raising against Gibbard's view. It is, I would say, a virtue of his view that it makes room for people to adopt stance (b), given that his aim is to be neutral regarding substantive moral questions such as whether moral reasons always override non-moral reasons. Nonetheless, the cases I have been developing demonstrate that it can be appropriate to adopt stance (c), and that doing so involves disagreeing with someone who adopts stance (a), and it is this disagreement that Gibbard is unable to account for as a disagreement in plans even after appealing to the norm-expressivist analysis of morality in his earlier work. As a result, either he must deny that there is a genuine disagreement

between Xanti and Plato, or between Portis and those of us that deem Mortis' murdering wrong even though we plan to murder in M and not to feel guilt in M, or else he must conclude that not all normative judgments can be captured by his view.

IV. CONCLUSION

If I have indeed identified a set of normative judgments that Gibbard's plan-based expressivism is unable to capture, what is the significance of that for the overall expressivist project? One possibility is that expressivists could simply reject Gibbard's latest formulation of the view in an attempt to avoid this problem. I am not prepared to rule out this possibility, but I think there is some reason to doubt that an alternative expressivist analysis will be able to avoid the problem I have identified without running afoul of other, perhaps even more serious, objections.

One obvious alternative is Gibbard's own earlier expressivist view, as formulated in *Wise Choices, Apt Feelings*. I've already indicated, however, why I think an appeal to norm-expressivism does not provide a way for an expressivist to capture all of the normative judgments involved in assessing an innocent mistake of mental constitution. Rather than work through a series of other forms of expressivism that have been or could be advanced, let me instead explain why I am skeptical that they will overcome this obstacle.

As Gibbard explains when motivating his project in *Thinking How to Live*, expressivism is particularly appropriate for understanding claims about what to do in a given situation because of the intimate connection between conclusions about what to do and action.[32] Gibbard's view is that any non-expressivist analysis of the thing to do is suspect because it threatens to sever this tie. I take this to be a powerful line of argument, and one that lies at the heart of any expressivist project. The beauty of Gibbard's plan-based expressivism is that it captures this idea more clearly than any other form of expressivism I have encountered. The cost of capturing this central idea so well, however, is that the resulting analysis reveals not only the strengths of expressivism, but also its limitations. What I have tried to illustrate with the examples involving innocent mistakes of mental constitution, either in the case of an ordinary moral agent evaluating Mortis or in the case of Xanti evaluating Soc, is that there are at least some genuine normative judgments that are not tied to action. In other words, the cases are designed to highlight normative judgments regarding an action in a situation that do not impinge on one's assessment of whether to perform the action in that situation, whether to plan to perform the action, whether to resent those who perform it, whether to feel guilt over performing it, or whether to have any other reactions. This is why they do not show up in any of our planning judgments, and why Gibbard's plan-based expressivism is incapable of capturing them. By extension, any view that begins with a commitment to capturing the intimate connection

between ordinary normative judgments and action can be expected to have a similar blind spot.

Perhaps a form of expressivism could be developed that would "see" these judgments and incorporate them into the expressivist analysis. That could only be done, however, at the cost of giving up the expressivist commitment to providing an account of the tie between normative judgment and action, which is to say giving up the primary motivation for offering an expressivist analysis in the first place. This would eliminate one of the chief comparative advantages claimed on behalf of expressivism, and make it much more difficult to defend expressivism against rival metaethical positions.

I would not rule out the possibility of devising a defensible form of expressivism that captures the normative judgments that escape Gibbard's plan-based analysis, but I think the more plausible alternative is to imagine a less ambitious form of expressivism. For all I have said here, Gibbard's analysis in *Thinking How to Live* may still do an excellent job of accounting for the vast majority of our normative judgments. And in fact I think it does. A more modest campaign for expressivism might content itself with that, and give up on the hope to provide an expressivist analysis of all normative judgments. Gibbard himself may not find this limited form of expressivism attractive, given his larger interest in extending the boundaries of expressivism, but a successful expressivist analysis

might have to be openly conscious of its limits and recognize the existence of genuine normative judgments that require some sort of descriptive analysis.

[*] For insightful comments on earlier versions of this paper, I am grateful to Stephen Darwall, Brian Epstein, Bill FitzPatrick, Allan Gibbard, Simon May, Nishi Shah, and an audience at the University of Southern California. Special thanks for their helpful comments and criticisms also go to Associate Editor Donald Hubin and several anonymous reviewers and editors of *Ethics*.

[1] See, for instance, A. J. Ayer, *Language, Truth, and Logic* (London: Gollantz, 1946), Charles Stevenson, *Ethics and Language* (New Haven: Yale University Press, 1944), R. M. Hare, *Moral Thinking* (New York: Oxford University Press, 1981), Simon Blackburn, *Spreading the Word* (Oxford: Oxford University Press/Clarendon, 1984), and Allan Gibbard, *Wise Choices, Apt Feelings* (Cambridge, MA: Harvard University Press, 1990).

[2] For a statement of this problem, see Peter Geach, "Assertion," *Philosophical Review* 74 (1965): 449-465.

[3] Ronald Dworkin, "Objectivity and Truth: You'd Better Believe It," *Philosophy and Public Affairs* 25 (1996): 87-139.

[4] Gibbard, *Thinking How to Live* (Cambridge, MA: Harvard University Press, 2003).

[5] When I speak of innocent mistakes I am not taking a stand on whether the agent is blameworthy for getting into the relevant situation. Rather, I use the term only to signify that the agent performs an action that is in some sense endorsed and in another sense condemned. As will become clear, the various types of innocent mistakes I discuss differ in the forms of endorsement and condemnation involved.

[6] For simplicity, I will follow Gibbard and focus on questions of what to do, rather than what to think or what to feel.

[7] This paragraph summarizes Gibbard's analysis in *Thinking How to Live*, pp. 48-53 and 68-71.

[8] Notice that Gibbard's contingency plans involve the sort of extreme perspective shift incorporated into Harsanyi's view of the original position. Rawls criticizes Harsanyi by denying that a chooser who imagines herself adopting the complete preferences and mental constitution of each relevant individual will be able to unify these incongruent perspectives and generate a meaningful assessment of expected utility. Such criticism may be apt with respect to Harsanyi's version of the original position, but for Gibbard the formation of a plan for any given situation involves the hypothetical adoption of just one perspective, that of the agent in the situation in question, and so the extreme perspective shift involved in hypothetical planning does not yield any disunity that could potentially impede plan-formation in the way that Rawls believes expectation-formation is impeded in the case of Harsanyi's original position. Thanks to an anonymous reviewer for identifying the parallel here. See John C. Harsanyi, "Morality and the Theory of Rational Behavior" *Social Research* 44 (1977): 623-656 and John Rawls, *A Theory of Justice: Revised Edition* (Cambridge, MA: Harvard University Press, 1999), at p. 150.

[9] The phrase "endorse your plan" is potentially ambiguous. To avoid the need for excessively convoluted alternatives, let me stipulate that if I endorse your plan that means that I endorse your planning the way you do, not that I endorse what you plan to do.

[10] The situation and notation presented here are taken from Gibbard, *Thinking How to Live*, 241-243.

[11] In invoking idealization I am following Gibbard's lead. Insofar as this use of idealization is objectionable, that constitutes an objection to Gibbard distinct from the arguments I am developing. Nonetheless, it is worth noting that at least some potential objections against idealization are inapplicable here. For instance, Gibbard is immune from Enoch's criticism of the use of idealization in response-dependence views because Gibbard is not making the claim that the judgments of an idealized agent are what *make* an action the thing to do. Rather, Gibbard's aim is

to be able to endorse everything a moral realist might want to say about what makes an action the thing to do by interpreting those claims expressivistically. This, perhaps, leads to a general concern about the objectivity of expressivism, along the lines of Dworkin's objections to expressivism. But such global concerns about objectivity are difficult to pin on expressivism, perhaps because expressivists have adequate responses available. In any case, my aim is to identify precise, localized areas of normative judgment in which expressivism lacks adequate objectivity. For Enoch's criticism of idealization, which I think Enoch would agree does not apply in this case, see David Enoch, "Why Idealize?" *Ethics* 115 (2005): 759-787.

[12] Notice that even though it is Xanti whose plans for what to do and what to plan to do conflict, in adopting these conflicting plans she is attributing a pure planning innocent mistake to Soc.

[13] Notice that the point here is not that Xanti cannot plan to plan to flee if in S, but simply that she cannot arrive at a plan to flee in S by following the strategy "defer to an idealized version of my actual self."

[14] This example should work for ε values ranging from complete deference (1) to total disregard (0).

[15] See Gibbard, *Wise Choices, Apt Feelings*, 204-217. Gibbard identifies this section of *Wise Choices, Apt Feelings* as addressing the possibility of what I am calling pure planning innocent mistakes in *Thinking How To Live*, p. 242, n. 4.

[16] Gibbard, *Wise Choices, Apt Feelings*, 213.

[17] Ibid., 215.

[18] Ibid., 217.

[19] There are, of course, cases of akrasia in which something like this occurs, but notice that attributing a pure planning innocent mistake involves seeing this dissonance between plans and action as supported by reasons rather than as akratic.

[20] Thanks to Donald Hubin for suggesting this line of objection.

[21] Gregory S. Kavka, "The Toxin Puzzle," *Analysis* 43 (1983): 33-36.

[22] Kavka rules out any resort to external incentives that will serve as reasons to drink the toxin, such as hiring a hit man to kill you if you don't drink it.

[23] For a survey of the literature on the toxin puzzle as well as the more general question of whether reasons to intend can diverge from reasons to carry out the intended action, see R. Clarke, "Commanding Intentions and Prize-Winning Decisions," *Philosophical Studies* 133 (2007): 391-409. Clarke argues that there can be reasons to intend to do something that are not also reasons to do it, but his examples cannot be used to support pure planning innocent mistakes because they are specifically designed such that the agent nonetheless has overall reason to perform the intended action.

[24] This creates a puzzle in the toxin case because a self-consciously rational agent cannot believe she will drink the toxin unless she believes she will have overall reason to do so, and the prize alone seems not to provide any such reason.

[25] Gibbard, *Thinking How to Live*, 260.

[26] Ibid., 269.

[27] Ibid., 239. See also Gibbard's objection to Hare on the grounds that Hare improperly assumes agreement between what Hare calls archangels. The central point there as well is that things like innocent mistakes of mental constitution are possible, and that an adequate theory cannot be based on the assumption that they will not occur. See Gibbard, "Hare's Analysis of 'Ought' and Its Implications," in *Hare and Critics*, ed. Douglas Seanor and Nicholas Fotion (Oxford: Oxford University Press/Clarendon, 1988) 57-72.

[28] Gibbard, *Thinking How to Live*, 243-248.

[29] This parallels the example in Section I of Alice and Jeremy failing to disagree about whether to take a drink.

[30] See especially Gibbard, *Wise Choices, Apt Feelings*, 36-54. Thanks to an anonymous reviewer and the editors of *Ethics* for pressing me on this point.

[31] Although this discussion follows Gibbard's lead from *Wise Choices* by focusing on resentment, it should be clear that any attempt to capture the disagreement in terms of plans for how to respond or react to Mortis, say by attempting to stop him or punish him (or whatever), will fail for similar reasons.

[32] Gibbard, *Thinking How to Live*, 8-13.