Figuring Out How to Proceed with Evaluation After Figuring Out What Matters

CHRISOULA ANDREOU *University of Utah*

ABSTRACT: I focus on David Gauthier's intriguing suggestion that actions are not to be evaluated directly but via an evaluation of deliberative procedures. I argue that this suggestion is misleading, since even the most direct evaluation of (intentional) actions involves the evaluation of different ways of deliberating about what to do. Relatedly, a complete picture of what an agent is or might be (intentionally) doing cannot be disentangled from a complete picture of how s/he is or might be deliberating. A more viable contrast concerns whether actions and *deliberative procedures are properly evaluated* on the whole *or, instead,* through time.

RÉSUMÉ : Dans cet article, je concentre mon attention sur l'intrigante suggestion de David Gauthier voulant que les actions ne doivent pas être évaluées directement, mais par le biais d'une évaluation des procédures délibératives. Je soutiens qu'il s'agit d'une fausse piste, car même l'évaluation la plus directe d'actions (intentionnelles) implique l'évaluation de différentes façons de délibérer sur la conduite à tenir. De façon connexe, on ne peut dresser un portrait complet de ce qu'un agent fait ou pourrait faire (intentionnellement) en faisant abstraction du portrait complet de la façon dont il ou elle délibère ou pourrait délibérer. Il est plus viable de se demander si les actions et *les procédures délibératives sont évaluées correctement* en entier *ou, plutôt,* à travers le temps.

Keywords: constrained maximization, David Gauthier, deliberative framework, deliberative procedure, intentional action, rationality, toxin puzzle

***Dialogue* 55 (2016), 621–637.**

doi:10.1017/S0012217316000548

I.

Traditional rational choice theory is founded on the assumption that what it is rational for an agent to do in a certain situation is determined by the agent's (ultimate) ends. In his pioneering work developing and defending the theory of rationality as constrained maximization, David Gauthier accepts this assumption, but rejects the closely related traditional assumption that the rationality of an action is *directly* determined by whether the action best serves the agent's ends.[1] Instead, Gauthier suggests that the rationality of an action is determined by whether the action is supported by a *deliberative procedure* (or system of deliberative procedures) that best serves the agent's ends.[2] In Gauthier's words, "the truth … about whether actions are rational or not … is settled by relating actions to deliberation, and the truth about the rationality of deliberative procedures is settled by determining which ones will prove most conducive to the agent's aim."[3] Gauthier's position has developed over time and, indeed, recently, Gauthier has made a move to replace the 'constrained maximizers' in his conception of deliberative rationality with 'rational cooperators' or 'agreed Pareto-optimizers.'[4] But Gauthier's preceding challenge regarding the nature of rationality, and, in particular, his suggestion that, roughly put, the evaluation of actions should proceed via the evaluation of deliberative procedures warrants further exploration.[5]

It might be supposed that the deliberative procedure that best serves an agent's ends is simply the deliberative procedure that tells the agent to always choose the action that best serves her ends. But, in his work on constrained maximization, Gauthier argues that this is not so. According to Gauthier, agents (or at least agents who are 'translucent' in that they cannot rely on being able to hide their true intentions) are better served by a deliberative procedure that sometimes calls for constraining behaviour in accordance with prior intentions, such as those that figure in prior sincere assurances. Agents with a deliberative procedure of this sort will be capable of making credible assurances to act in ways that will require them to show constraint. This capacity is very important, since benefits are sometimes dependent on being able to provide such credible assurances. For example, even if cooperating in harvesting our crops would be mutually beneficial, you may not agree to help me with my crops if my crops have to be harvested first and my assurance to help you if you help me is recognizably insincere. To get my translucency facilitating, rather than impeding, mutually beneficial cooperation, I need to be able to provide a

---

1 Gauthier 1994.

2 Gauthier 1994, pp. 701–702.

3 Gauthier 1994, p. 702.

4 Gauthier 2013.

5 Except where I indicate otherwise, attributions to Gauthier are based on his position in "Assure and Threaten," 1994, where he both builds on and makes some important revisions to his position in *Morals by Agreement*, 1986.

sincere assurance to help you if you help me; and to do that, I need to have a deliberative procedure that calls for constraining my behaviour in accordance with certain prior intentions, like the intention that figures as part of the sincere assurance to help you harvest your crops if you help me harvest mine. Given that benefits are sometimes dependent on being able to provide credible assurances that call for future constraint, the deliberative procedure that tells the agent to always choose the action that best serves her ends cannot be the best deliberative procedure—agents fare better with a deliberative procedure (or system of procedures) that at least sometimes calls for genuine constraint.[6]

Here is another way of thinking about the contrast between traditional rational choice theory and Gauthier's theory of rationality as constrained maximization: While the theories agree about the *primary criterion* of evaluation, in that they both take the promotion of the agent's ends to be what matters, they disagree about the *primary objects* of evaluation, understood as the objects to which the primary criterion of evaluation should be directly applied. Traditional rational choice theory casts individual actions (or 'moves' that can be made at individual 'decision points') as the primary objects of evaluation, while Gauthier's theory of rationality as constrained maximization casts deliberative procedures as the primary objects of evaluation. Taking individual actions as the primary objects of evaluation commits one to judging deliberative procedures in terms of their fit with the criterion-favoured action(s). Taking deliberative procedures as the primary objects of evaluation commits one to judging individual actions in terms of their fit with the criterion-favoured deliberative procedure(s). Either way, at any choice point, one is ensured a coordinated action-and-deliberative-procedure package that determines what counts as choosing well; but the package that emerges depends, it seems, on whether individual actions or, instead, deliberative procedures are taken as the primary objects of evaluation.[7] Gauthier's challenge, described in terms of the

---

[6] Note that, as far as I can tell, Gauthier's reasoning is consistent with the idea that the question of what matters and the question of what deliberative procedure (or procedures) is (or are) correct (given what matters) are 'theoretical' questions, in the sense that the answers are truths that can be 'discovered.' Indeed, one might interpret Gauthier as defending the following as (objective) truths: (i) what matters is serving one's ends well; and (ii) a deliberative procedure is correct if employing it promotes what matters as well as possible (i.e., best). Discussion of this issue and of David Velleman's related critique in "Deciding How to Decide," 2000, are beyond the scope of this paper.

[7] This way of casting Gauthier's theory is, in some ways, similar to Michael Thompson's in his work on practices and dispositions (see Thompson 2008, III.10), but my approach highlights the fact that a 'transfer principle' of some sort is involved whatever objects of evaluation are put forward as primary. Also, I focus on the 'rational deliberative procedures version' of Gauthier's theory (see, especially, Gauthier 1994) rather than on the 'rational dispositions version' (which figures in *Morals by Agreement*, 1986).

idea of primary objects of evaluation, is that, if individual actions are taken as the primary objects of evaluation, then rationality emerges as an obstacle to providing beneficial sincere assurances in cases such as the harvesting case above, whereas, if deliberative procedures are taken as the primary objects of evaluation, then rationality emerges as consistent with showing, and achieving the benefits of, genuine constraint. With this in mind, Gauthier maintains that an agent's primary concern should be with "the overall effect of employing certain deliberative procedures," taking into account "not only the actions they determine, but also the actions they make possible."[8]

Now plenty of philosophers resist the idea, accepted by both Gauthier and traditional rational choice theorists, that what it is rational for an agent to do in a certain situation is determined simply by the agent's ends. Many hold that being fully rational involves being sensitive to moral reasons for action. Gauthier (or at least the time-slice of Gauthier I am here concerned with) argues that these two seemingly conflicting ideas can, to some extent, be reconciled, and that the latter idea can and should be grounded in the former via his suggestion that deliberative procedures are the primary objects of evaluation. For Gauthier, a compelling conception of morality must combine the following two features: it must represent morality as requiring that agents show genuine constraint rather than always performing the actions that best serve their ends; and it must ground morality in an agent-centred theory of rationality. Gauthier's focus on deliberative procedures rather than individual actions is presented as the key to satisfying this seemingly inconsistent pair of requirements. According to Gauthier, being moral essentially involves accepting the requirement to genuinely constrain oneself on the basis of prior assurances that one benefited from providing, as per the demands of rational deliberation, understood as deliberation employing a deliberative procedure (or system of procedures) that best serves the agent's ends.

But, even if one rejects Gauthier's idea that the rational authority of morality is properly grounded in an agent-centred theory of rationality, there is still reason to dwell on the possibility, highlighted by Gauthier's position, that, with regards to practical rationality, there can be agreement about the primary criterion of evaluation without there being agreement about the primary objects of evaluation.[9] This possibility may be of crucial significance, regardless of whether commitment to an agent-centred theory of rationality is warranted. Even for a non-agent-centred criterion of evaluation, say the promotion of K, there can be agreement that the promotion of K is what matters (either invariably or

[8] Gauthier 1994, p. 701.

[9] As Michael Thompson notes, while the same sort of possibility with regards to morality is highlighted by practice versions of utilitarianism, it is to Gauthier especially that we owe the idea of a similarly structured theory of rationality. See Thompson 2008, p. 149 and p. 167.

else just in a certain situation), but disagreement about whether it is individual moves or deliberative procedures that are to be selected on the basis of this criterion.

My initial focus in this paper will be on the idea that individual moves should, at least sometimes, be evaluated in terms of their connection with a more primary object of evaluation. My sense is that there is something to this view, and that its merits can be disentangled from debates about whether we should accept an agent-centred theory of rationality. Still, as will become apparent, I think it is misleading to suggest that (past, current, or potential future) intentional actions (which are the actions of interest here) are not to be evaluated directly but via an evaluation of deliberative procedures. In my view, even the most direct evaluation of intentional actions involves the evaluation of different ways of deliberating about what to do. Relatedly, a complete picture of what an agent is or might be (intentionally) doing cannot be disentangled from a complete picture of how he is or might be deliberating. I will end by suggesting that the task of figuring out how to proceed with evaluation after figuring out what matters is still very much a pressing task, but the question that needs to be answered is not the question of whether actions or deliberative procedures are the primary objects of evaluation; rather, it is the question of whether actions *and* deliberative procedures and frameworks are to be evaluated *on the whole* or, instead, *through time*.

II.

How might one defend the view that individual moves should, at least sometimes, be evaluated in terms of their connection with a more primary object of evaluation? Consider the following suggestion: when an individual move is chosen as part of a 'larger' action, the acceptability of the move can depend on the acceptability of the larger action embarked on / underway.[10] Suppose, for example, that Alice faces three options: [1] she can buy item X at store S; [2] she can incur an extra cost to go to store S^-, which is somewhat out of the way, and buy X^+, an item that is somewhat better than X; or [3] she can incur an extra cost to go to S^- and buy X^-, an item that is somewhat worse than X. Suppose that, given the locations of S and S^-, and the differences between and X, X^+, and X^-, options [1] and [2] are equally good, while option [3] is bad (relative to whatever it is that matters, whether that is the agent's ultimate ends or something else altogether). Now suppose we see Alice leaving store S. Is this a bad move? It seems quite plausible to answer that it is not a bad move if Alice is going to store S^- to buy X^+, but it is a bad move if Alice is going to S^- to buy X^-. But this is basically to say that whether Alice's leaving store S is a bad move depends on what larger action is underway.

[10] This suggestion figures crucially in my articles Andreou 2014 and Andreou 2006a.

https://doi.org/10.1017/S0012217316000548 Published online by Cambridge University Press

One can perhaps resist this conclusion via acceptance of the following picture: The move of leaving S is not a bad move even if Alice is going to S^- to buy X^-. A bad move occurs only when Alice buys X^- rather than X^+. The initial move sounds bad if it is described as leaving S to go to S^- and buy X^-, but this description goes beyond describing Alice's initial move, which is simply leaving S. Note, however, that, even if one counts leaving S as an acceptable move, this is presumably because leaving S *could* figure as part of an acceptable larger action, namely leaving S to go to S^- to buy X^+, so the move of leaving S is still being evaluated in terms of a larger action—it is just that the larger action is merely one that *could* be underway rather than one that *is* underway. The primary object of evaluation is still a larger action. If one abstracts not only from the action embarked on in leaving S, but even from the larger actions that could be embarked on via leaving S, it seems impossible to evaluate the move of leaving S at all; one could say that leaving S involves incurring a cost, but could not say anything about whether leaving S is worth the associated cost; the move must be evaluated in terms of its connection with a more primary object of evaluation.

Now let us grant that there is a notion of (minimal) rationality according to which leaving S is rational (or rationally permissible) because (or in the sense that) leaving S *could* be part of the acceptable action of leaving S to go to S^- to buy X^+. It still seems undeniable that, if Alice's leaving S is *actually* guided by the goal of going to S^- to buy X^- rather than the goal of going to S^- to buy X^+, then some sort of negative evaluation of Alice's move is in order, since the move is guided by a goal that is, by hypothesis, unacceptable. We can capture this negative evaluation by saying that while Alice is leaving S, her behaviour is 'misguided.'[11]

It might be objected that, even in the scenario in which Alice has the goal of buying X^-, her behaviour, when she is leaving S, is not yet *guided* by the goal of buying X^-; it is, as yet, guided only by the permissible goal of leaving S. But this is false. Notice first that, while leaving S, Alice is already embarked on going to S^-. In this early stage of going to S^- (just as in later stages and in any preparatory stages), Alice is guided by the goal of getting to S^-, and she would, quite possibly, not be leaving S at all were she not going to S^-. Relatedly, while leaving S, Alice is already embarked on going to buy X^-. In this early stage of going to buy X^- (just as in later stages and in any preparatory stages), Alice is guided by the goal of buying X^-, and she would, quite possibly, not be leaving S at all apart from her (let us suppose) correct belief that she can buy X^- at S^-, a belief that would be

[11] The distinction in this paragraph is related to Anita Superson's distinction in *The Moral Skeptic* between assessing "bare acts" and assessing "actions as performed" by the agent. See Superson 2009, p. 181.

https://doi.org/10.1017/S0012217316000548 Published online by Cambridge University Press

irrelevant if her behaviour were guided simply by the goal of leaving S.[12] Since Alice's leaving S is guided by a problematic goal, it really is misguided.

For our purposes, what matters is that the evaluation of an individual move (as rational or irrational, well-guided or misguided) will at least sometimes depend on the evaluation of either an actual or a possible larger action. Note that I will not distinguish sharply between (temporally extended) actions and courses of action. As will become clear in the remaining sections, I see the phases of an action or course of action as united by the deliberative constraints that would have to be in play for each phase to count as directed toward the same end or object.

III.

Turn now to my suggestion that even the most direct evaluation of intentional actions involves the evaluation of different ways of deliberating about what to do. Consider the possibility of intentionally Z-ing, where Z-ing is an action (or a course of action) that figures as one among a small set of equally good options. For an agent to count as intentionally Z-ing, the agent's deliberative framework must have certain features. For instance, suppose an agent believes that Y-ing is necessary for Z-ing. If the consideration that Y-ing is necessary for Z-ing does not settle, for the agent, the question of whether to Y, then the agent does not count as intentionally Z-ing.[13] Suppose, for example, an agent (correctly) believes that, to go to his office, he must turn left at the upcoming light. If the consideration that he must turn left to go to his office does not settle for him the question of whether to turn left at the upcoming light, then he is not (intentionally) going to his office. Perhaps, like Buridan's Ass,[14] he is paralyzed by the choice he faces and so has not even pulled out of his driveway. In any case, even if he is moving, he does not count as intentionally going to his office until he has incorporated some relevant constraints into his deliberative framework.[15]

As the case just described suggests, an agent's deliberative framework normally changes over time, with frameworks expiring if and when associated intentional actions are completed. If, on Monday morning, an agent is intentionally going to his office and believes that to go to his office he must turn left

[12] Certain features of my description of this case pick up on some points concerning actions in progress emphasized in Thompson 2008, II.

[13] I defend very closely related points concerning intentions, using illustrations similar to those incorporated below, in Andreou 2009 and Andreou 2006b.

[14] In a simple version of this hypothetical dilemma, an ass is placed right in the middle of two identical piles of hay and, with no basis for choosing one pile over the other, it starves to death.

[15] These points about intentionally doing something are closely related to points in Bratman 1987 concerning how intentions frame one's future deliberations.

at the upcoming light, then the consideration that he must turn left to go to his office settles for him the question of whether to turn left. But, on Monday night, when the agent is, let us suppose, intentionally going to the grocery store, the consideration that he must turn left to go to his office may not play any important role in his deliberative framework. Still, much of an agent's deliberative framework can remain stable over time. If, for example, an agent values being a good friend and believes that (in normal circumstances) being totally honest is necessary for being a good friend, then it may be a stable feature of the agent's deliberative framework that the consideration that being totally honest is necessary for being a good friend settles for her the question of whether to be totally honest with her friends when (in normal circumstances) they ask her for advice or feedback. Moreover, if being a good friend really does require being totally honest when (in normal circumstances) a friend asks for advice or feedback, then so long as A is a good friend to B, it is a stable feature of A's deliberative framework that (in normal circumstances) a consideration of the form 'it is necessary for being totally honest with B to tell B that P' settles for A the question of whether to tell B that P. To take a somewhat different example, if an agent values being reliable and believes that factoring in some time for unexpected delays is necessary for being reliable, then it may be a stable feature of the agent's deliberative framework that the consideration that factoring in some time for unexpected delays is necessary for being reliable settles for her the question of whether to factor in some time for unexpected delays when making commitments and constructing her schedule. Moreover, if being reliable really does require factoring in some time for unexpected delays, then, so long as A is reliable, it is a stable feature of A's deliberative framework that a consideration of the form 'allotting at least N amount of time to task T is necessary given the possibility of unexpected delays' settles for A the question of whether to allot at least N amount of time to task T if she commits to the task.

The (relatively) stable core of an agent's deliberative framework can include elements from the very abstract deliberative procedures with which Gauthier is concerned. One might, for example, take considerations of the form 'action X will best serve my ends' as settling the question of whether to X. One can do this while keeping in mind that (i) whether an action best serves one's ends may depend on what larger (course of) action it is part of, and that (ii) to intentionally X is to incorporate additional constraints into one's deliberative framework. Alternatively, one might, like Gauthier, favour a more constrained core and take considerations of the form 'V-ing is necessary for following through on a prior intention that I benefited from forming' as settling the question of whether to V.

Return now to the idea that intentionally Z-ing involves incorporating certain constraints into one's deliberative framework. It follows from this that directly evaluating the rationality of intentionally Z-ing cannot be disentangled from evaluating the rationality of incorporating certain constraints into one's

deliberative framework. Suppose, for example, that, given his skills, interests, and the means involved, it would be rational for Angelo to go to his office or to go to the movie house—either action is acceptable, though he cannot do both. If he decides to go to his office, then, while his going is underway (i.e., while he is intentionally going to his office), his deliberative framework is adjusted so that, for example, the consideration 'turning left at the upcoming light is necessary for going to my office' settles the question of which direction to turn at the upcoming light. If he decides to go to the movie house, then, while his going is underway (i.e., while he is intentionally going to the movie house), his deliberative framework is adjusted so that, for example, the consideration 'turning right at the upcoming light is necessary for going to the movie house' settles the question of which direction to turn at the upcoming light. To accept that, given his skills, interests, and the means involved, it would be rational for Angelo to go to his office or to go to the movie house (but not both) is to accept that it would be rational for Angelo to deliberate in accordance with either (but not both) of the adjusted frameworks just described.

Given that directly evaluating the rationality of an (intentional) action involves evaluating an associated deliberative framework, and given that, in at least some cases, the rationality of an agent's move should or perhaps even must be evaluated in terms of the rationality of a larger (actual or possible intentional) action, it follows that, in at least some cases, the rationality of an agent's move should or perhaps even must be evaluated in terms of the rationality of a deliberative framework. To return to Angelo's case, Angelo's turning right is rational precisely because of its (potential) connection to his (intentionally) going to the movie house, which is inseparable from the deliberative framework that lets the consideration 'turning right is necessary for going to the movie house' settle the question of which direction to turn at the light. So the rationality of Angelo's move depends on the rationality of employing this deliberative framework.

This sounds like a somewhat Gauthierian conclusion. Notice, however, that there is no question here of applying the criterion of evaluation to deliberative frameworks *rather than* to actions. For having the relevant deliberative framework and (intentionally) performing the relevant action cannot be disentangled. Furthermore, evaluating the rationality of the move of turning right and evaluating the rationality of the action the agent is or could be embarked on cannot be disentangled. So, at least here, the situation is not one of having to choose between two conflicting verdicts concerning the rationality of a move, one resulting from casting actions as the primary objects of evaluation and the other resulting from casting deliberative frameworks as the primary objects of evaluation.

IV.

But what about cases in which the choice between two conflicting verdicts does present itself? Consider Gregory Kavka's toxin case: An agent will get a

https://doi.org/10.1017/S0012217316000548 Published online by Cambridge University Press

million dollars if he forms the intention to drink a toxin that will make him sick for a day.[16] He need not actually drink the toxin to get the million dollars. He gets the money if, and only if, tonight, he intends to drink the toxin when it becomes available tomorrow. In this case, it does seem that one can directly evaluate drinking the toxin and get a different verdict about the rationality of drinking than if one were to evaluate drinking indirectly via an evaluation of the deliberative framework or procedure that calls for maximizing modulo sticking to prior intentions that one benefited from forming. Borrowing from Gauthier (and abstracting from some complications that need not concern us here), let us call this 'the procedure of constrained maximization.' Drinking the toxin does not seem to promote what matters (which, in the case at hand, is, by hypothesis, mainly financial gain), since nothing is to be gained by drinking the toxin; but having the deliberative procedure of constrained maximization, which calls for drinking the toxin after forming the intention to drink it, does seem to promote what matters, since it allows one to form the intention to drink the toxin and thus get the million dollar reward. Should one take the drinking of the toxin or the deliberative procedure of constrained maximization as the primary object of evaluation?

Let us back up for a moment. Does having the deliberative procedure of constrained maximization really serve the agent well (which, let us assume, is all that is at stake)? Suppose the potential toxin-drinker raises this question *when it is time to drink or refrain from drinking the toxin*. He has, let us assume, reasoned as a constrained maximizer to date (and so has earned the million dollar reward), but is now somehow prompted to step back and reflect on this question. And now consider two possibilities, keeping in mind Gauthier's idea that "the fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection."[17]

Suppose that a switch in deliberative procedures from constrained maximization to (the traditionally favoured procedure of) straightforward maximization, even if possible and even if it would have no negative reputation effects, would still be disadvantageous because the potential toxin drinker would miss or risk missing opportunities available only to constrained maximizers. Perhaps, even if his switch to straightforward maximization is not permanent and even if he can creatively reinterpret his lapse later, during his lapse, he is likely to miss out on one or more extremely beneficial offers—not because of any reputation effects but because of his translucency.[18] Suppose the potential toxin drinker,

16 Kavka 1983.

17 Gauthier 1986, p. 183.

18 As Gauthier's work suggests, one can lose out not only by being unable to take advantage of certain potentially beneficial offers (due to one's translucency), but also by failing to even receive certain potentially beneficial offers (due to one's translucency).

impressed by the ongoing advantages of constrained maximization, reaffirms his confidence in the rationality of the procedure, and, remaining resolute (by choice rather than due to inertia), proceeds to drink the toxin because he benefited from forming the intention to do so. In this scenario, the individual move of drinking the toxin seems to figure as an essential part of a larger, and in this case intended, (course of) action, namely (the course of) remaining resolute. By remaining resolute, the toxin drinker avoids any lapses in beneficial opportunities available only to constrained maximizers. But, as part of remaining resolute, he takes considerations like 'drinking the toxin is necessary for sticking to a prior intention I benefited from forming' as settling the question of whether to drink the toxin. This version of the toxin case seems very much like the driving case above, in that the rationality of a certain move cannot be disentangled from the rationality of a larger (course of) action (actually or potentially) underway, and the larger action cannot be disentangled from an associated deliberative framework or procedure. So, in this scenario, the toxin case does not present us with a situation in which we face and must choose between two conflicting verdicts concerning the rationality of a move, one resulting from casting actions as the primary objects of evaluation and the other resulting from casting deliberative procedures as the primary objects of evaluation.

Gauthier might object that, even though an agent can, in light of reflection, decide to remain resolute (or decide not to), remaining resolute is not an action or course of action—it describes a state of mind. But there is a false contrast here. Like run-of-the-mill cases of intentionally Z-ing, such as going to one's office (which has a built-in termination point) or jumping rope (which does not), as well as more far-reaching cases of intentionally Z-ing, such as earning one's Ph.D. or keeping a long-distance friendship alive, remaining resolute involves having a certain deliberative framework and acting (or at least attempting to act) accordingly.

Moving on to the next version of the toxin case, suppose, alternatively, that, when the time comes to drink or refrain from drinking the toxin, it is somehow completely clear that a switch from constrained maximization to straightforward maximization, if possible, would be advantageous, and that this is so taking into account not only reputation effects—including the effects of one's behaviour on one's *self*-image, which can have further effects—but also any opportunities that might come up for one were one to remain a constrained maximizer. (If switching deliberative procedures could never be advantageous, then there is no alternative scenario to consider and my argument is done.) Perhaps it has become clear that, if one remains a constrained maximizer, one will die before any new opportunities dependent on one's being a constrained maximizer would come knocking but not before one will have to incur the cost of remaining a constrained maximizer, namely, being sick for a day after resolutely drinking the toxin. Could the potential toxin drinker still reaffirm the rationality of constrained maximization? Is constrained maximization still the deliberative procedure that best serves the agent's ends?

It might seem like the correct response must be 'no.' Given that, by hypothesis, switching deliberative procedures from constrained maximization to straightforward maximization would be advantageous, even given the agent's translucency, constrained maximization no longer best serves the agent's ends, even if it once did.[19] If this line of reasoning is correct, then again the toxin case does not present us with a situation in which we face and must choose between two conflicting verdicts concerning the rationality of a move, one resulting from casting actions as the primary objects of evaluation and the other resulting from casting deliberative procedures as the primary objects of evaluation. Rather, when the time comes to drink or refrain from drinking the toxin, the deliberative procedure that best serves the agent's ends is straightforward maximization and it calls for refusing to drink the toxin, which is the action that best serves the agent's ends.

One might insist, however, that the best deliberative procedure (among a set of mutually exclusive alternatives) cannot vary over time because (distracting qualifications aside) the best deliberative procedure should be understood as the procedure that, employed consistently over time, best serves the agent's ends; and it is always true of constrained maximization that, employed consistently over time, it serves the agent's ends better than any alternative deliberative procedure employed consistently over time. Here 'best' can be glossed as 'best on the whole,' where a deliberative procedure can count as best on the whole even if one will in no way benefit from having this deliberative procedure from time t on, not even via opportunities afforded to those with this procedure.

Let us make room for this approach and see where it leads us. Getting back to the version of the toxin case under consideration, we would have to say that, even though switching deliberative procedures from constrained maximization to straightforward maximization would be advantageous, the deliberative procedure of constrained maximization is still best on the whole. There is thus a conflict between the action that best serves the agent's ends—namely, refusing to drink the toxin—and the action called for by the best deliberative procedure—namely, drinking the toxin. But, as will become clear, this apparent rift in verdicts is due to mixing evaluative approaches: actions (or courses of action) are evaluated *through time*—with no block against the possibility of switching courses affecting the evaluation of an action (or course of action) as one progresses through its phases, even when everything proceeds as anticipated—whereas deliberative procedures are evaluated *on the whole*—with a built-in block against the possibility of switching procedures affecting the evaluation of a deliberative procedure as it is employed over time, at least when everything proceeds as anticipated. Here and below it is important to keep in mind

[19] This is a version of the 'reversion' problem Duncan MacIntosh raises in his work on Gauthier's position. See MacIntosh 1991.

that I do not distinguish sharply between (temporally extended) actions and courses of action, since I see the phases of an action or course of action as united by the deliberative constraints that would have to be in play for each phase to count as directed toward the same end or object. (Notice that typical (and perhaps all) temporally extended actions (including the ones I am concerned with here) have phases that allow the action to (also) be seen as a course of action.)

Note that moving from one (course of) action to another, or from one deliberative framework or procedure to another need not count as switching (in the sense I am interested in); in cases of switching, something is abandoned 'mid-stream,' before completion or expiry. For example, having a meal and then going for a walk does not, other things equal, count as switching courses, but putting some water on for tea and then turning it off to go deal with an emergency in the yard does count as switching courses. Similarly, moving from a deliberative framework that has expired (with the completion of the associated intentional action) to another framework does not, other things equal, count as switching frameworks, but moving from a framework that has not yet expired (or, like many deliberative procedures, has no built-in expiry) to an incompatible framework does count as switching frameworks. As assumed in my discussion above, because constrained maximization and straightforward maximization do not have built-in expiries, moving from one of these incompatible procedures to the other does count as switching frameworks.

Like the deliberative procedures we have been focusing on, and like deliberative frameworks generally, actions (or courses of action) can be evaluated on the whole rather than through time with the result that the (course of) action that is best on the whole can fail to serve the agent's concerns as well as a sequence of moves that amounts to switching courses over time. Consider, relatedly, Gauthier's idea that constrained maximization calls for "the best course of action *[the agent] can choose to follow*" (my emphasis).[20] The quoted phrase seems to leave room for the idea that the best course of action (full stop) might be one the agent *cannot* choose to follow, such as intending to drink the toxin at time t and then refusing to drink it when time t arrives. But, as Gauthier himself suggests in discussing the toxin case, intending to drink the toxin at time t and then refusing to drink it when time t arrives involves switching courses, rather than proceeding in a way that is unified enough to be "embrac[ed]" as a "single course of action."[21] The switch, as I see it, is from (the temporally extended action / course of action of) navigating oneself with an eye to drinking the toxin at time t to (the alternative of) navigating oneself with an eye to *not* drinking the toxin at time t. (I will say more about such switching below.) Given that a certain sequence of moves can fail to be unified

[20] Gauthier 1994, p. 695.

[21] Gauthier 1998, p. 48.

https://doi.org/10.1017/S0012217316000548 Published online by Cambridge University Press

enough to count as an action (or a course of action), there is room for counting an action (or a course of action) as best on the whole—in that it serves the agent's concerns better than any other (course of) action—even though switching courses would serve the agent's concerns even better.

Consider the situation in which the agent forms the intention to drink the toxin at time t and then drinks it when time t arrives. Throughout the time the agent intends to drink the toxin (preparing the way by, at a minimum, avoiding conflicting engagements), embarks on drinking it (lifting it to his lips), and follows through (swallowing the nasty stuff), the agent's deliberation is constrained by the same end or object—he is navigating himself with an eye to drinking the toxin at time t, and is thus pursuing the same (course of) action.[22] Whether time t has arrived or still lies ahead, the agent takes considerations of the form 'X-ing is necessary for drinking the toxin at time t' as settling the question of whether to X. This framework unites the phases of the agent's navigating himself with an eye to drinking the toxin at time t. This (course of) action can, like the procedure of constrained maximization, count as the (course of) action that is best on the whole even though switching courses and refusing to drink the toxin after forming the intention to drink it serves the agent's concerns better. Indeed, navigating oneself with an eye to drinking the toxin at time t will presumably count as the (course of) action that is best on the whole precisely when drinking the toxin at time t is called for by the procedure of constrained maximization; for, that is when navigating oneself with an eye to drinking the toxin at time t will be the best possibility that does not involve switching courses.

The point is that, while evaluation can be steered in different directions depending on whether one thinks of rationality in terms of evaluations of things on the whole or instead in terms of evaluations through time, once one of these perspectives is accepted as the correct one, evaluation does not seem to be steered in different directions depending on whether actions or, alternatively, deliberative procedures are taken as the primary objects of evaluation. Return to the second version of the toxin case under consideration, in which a switch from constrained maximization to straightforward maximization, if possible, would be advantageous. If actions and deliberative procedures are evaluated through time, with no block against the possibility of switching courses or procedures affecting evaluation (as specified above), then, when the time comes to drink or refrain from drinking the toxin, refraining from drinking the toxin and the procedure of straightforward maximization emerge as best. If, on the other hand, actions and deliberative procedures are evaluated on the whole, with a built-in block against the possibility of switching courses or procedures affecting evaluation, then, navigating oneself with an eye to drinking the toxin at time t and the procedure of constrained maximization emerge as best.

---

22 This point combines ideas from Bratman 1987 and Thompson 2008, II.

V.

My reasoning can be made clearer and, I hope, more compelling via consideration of the following toxin-like puzzle: Suppose an action theorist with a large grant offers an agent the following deal: she will get a million dollars if, in five minutes from now, she is intentionally going to her office. Suppose the trip would take 20 minutes and the agent knows that the first half of the trip would be no trouble but the second half of the trip would involve stressful driving through construction. The action theorist makes it clear that the agent will get the million dollars in five minutes if she is then going to her office and the money will not be revoked after that even if she never actually goes to her office (i.e., she never completes her action-in-progress). The rest of the details of the case are such that, by hypothesis, the potential office-goer does best if, in five minutes, she is going to her office but, soon thereafter, she is giving up and turning back.

In this new toxin-like case, the million dollar reward is for intentionally doing something at a certain point in time. But, as I've been emphasizing, what an agent is intentionally doing at a certain point in time cannot be disentangled from her deliberative framework at that time. In particular, an agent does not count as intentionally going to her office just because she is heading in the direction of her office. Her deliberative framework must satisfy certain constraints. For example, if she believes that to go to her office she must continue on this road for 15 more minutes, then she is not intentionally going to her office if she does not take that consideration as settling the question of whether to continue on this road for 15 more minutes.[23]

Now consider the question of whether the agent's concerns are best served by her going to her office. It is cryptic to say that the agent's concerns are best served by her going to her office and by her not going to her office. And it is misleading to say that the agent's concerns are best served by her heading in the direction of her office and then turning back, since the million dollar reward is for (being in a phase of) going to her office, not for heading in the direction of her office. Should one say that the agent's concerns are best served by her being in a phase of going to her office and then turning back? This involves saying that the agent's concerns are best served by her switching courses. And this is where we must decide whether to evaluate actions on the whole (so that the evaluation of an action is fixed regardless of one's temporal location relative to the action) or through time. If one evaluates actions on the whole, with a built-in block against the possibility of switching courses affecting evaluation, then one can say that, on the whole, the agent's concerns are

[23] This toxin-like puzzle and my discussion of it in this section are very much inspired by certain aspects of Michael Thompson's discussion of actions-in-progress in Thompson 2008, II. It also draws on my work on intentions in Andreou 2009 and Andreou 2006b.

https://doi.org/10.1017/S0012217316000548 Published online by Cambridge University Press

best served by her going to her office (rather than by her not going to her office). If one evaluates actions through time, with no block against the possibility of switching courses affecting evaluation, then, one can say that, before getting the million dollars, the agent's concerns are best served by her going to her office, but, after getting the million dollars, the agent's concerns are best served by her switching courses and turning back. Clearly, the question of which evaluative approach to take applies just as well to actions as to deliberative procedures, and consistency seems to require taking the same approach for both.

Return now to the original toxin case. Intending to drink the toxin at time t and then drinking it when time t arrives each figure as phases in navigating oneself with an eye to drinking the toxin at time t. These phases are united by features of the agent's deliberative framework in the same way the different phases of going to one's office are united by features of the agent's deliberative framework. If this is right, then there is no deep structural difference between the original toxin case and the toxin-like case I've presented; and, in the latter case, it is completely clear that, as with deliberative procedures, actions (or courses of action) can be evaluated on the whole or through time.

VI.

In his work on constrained maximization, Gauthier suggests that actions are to be evaluated indirectly, via an evaluation of deliberative procedure(s). In my view, this suggestion is misleading, since even the most direct evaluation of (intentional) actions involves the evaluation of different ways of deliberating about what to do. Relatedly, a complete picture of what an agent is or might be (intentionally) doing cannot be disentangled from a complete picture of how he is or might be deliberating. The task of figuring out how to proceed with evaluation after figuring out what matters remains, but the question that needs to be answered is not the question of whether actions or deliberative procedures are the primary objects of evaluation; rather, it is the question of whether actions *and* deliberative procedures are properly evaluated *on the whole* or, instead, *through time*.[24]

Acknowledgements: The development of this paper was greatly influenced by my participation in the "Rational Choice Contractarianism: 25 Years After *Morals by Agreement*" conference at York University. My thanks to all the organizers and participants. I am particularly grateful to Susan Dimock, who spearheaded the event, and to Michael Bratman, David Gauthier, Duncan MacIntosh, Anita Superson, and an anonymous referee, whose comments prompted substantial adjustments and additions. Thanks also to Elijah Millgram and Mike White for their very helpful written comments.

[24] Gauthier's work provides some thought-provoking ideas pertinent to debate redirected at this question; but that is a topic for another occasion.

References

Andreou, Chrisoula
2014 "The Good, the Bad, and the Trivial," *Philosophical Studies* 169 (2), 209–225.
Andreou, Chrisoula
2009 "Taking On Intentions," *Ratio* 22 (2), 157–169.
Andreou, Chrisoula
2006a "Temptation and Deliberation," *Philosophical Studies* 131 (3), 583–606.
Andreou, Chrisoula
2006b "Might Intentions Be the Only Source of Practical Imperatives," *Ethical Theory and Moral Practice* 9 (3), 311–325.
Bratman, Michael
1987 *Intention, Plans, and Practical Reason*, Cambridge: Harvard University Press.
Gauthier, David
2013 "Twenty-Five On," *Ethics* 123 (4), 601–624.
Gauthier, David
1998 "Rethinking the Toxin Puzzle," in J.L. Coleman and C.W. Morris (eds.), *Rational Commitment and Social Justice*, Cambridge: Cambridge University Press, 47–58.
Gauthier, David
1994 "Assure and Threaten," *Ethics* 104 (4), 690–721.
Gauthier, David
1986 *Morals by Agreement*, Oxford: Clarendon Press.
Kavka, Gregory
1983 "The Toxin Puzzle," *Analysis* 43, 33–36.
MacIntosh, Duncan
1991 "Preference's Progress," *Dialogue* 30 (1–2), 3–32.
Superson, Anita
2009 *The Moral Skeptic*, Oxford: Oxford University Press.
Thompson, Michael
2008 *Life and Action*, Cambridge: Harvard University Press.
Velleman, David
2000 "Deciding How to Decide," in *The Possibility of Practical Reason*, Oxford: Clarendon Press, 221–243.