DOI: 10.1111/phpr.13033

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Embracing self-defeat in normative theory

Samuel Fullhart

Princeton University, Princeton, New Jersey, USA

Correspondence
Samuel Fullhart, Princeton University, Princeton, New Jersey, USA.
Email: fullhart@princeton.edu

Abstract

Some normative theories are self-defeating. They tell us to respond to our situations in ways that bring about outcomes that are bad, given the aims of the theories, and which could have been avoided. Across a wide range of debates in ethics, decision theory, political philosophy, and formal epistemology, many philosophers treat the fact that a normative theory is self-defeating as sufficient grounds for rejecting it. I argue that this widespread and consequential assumption is false. In particular, I argue that a theory can be self-defeating and still internally consistent, action-guiding, and suitable as a standard for criticisn.

KEYWORDS

self-defeat, Prisoner's Dilemma, Professor Procrastinate, diachronic action, collective action

1 | INTRODUCTION

A group of us have come together to form a utopian community. We're all discussing ideas about how to live together going forward. We want to come up with a social scheme that will respect each member's autonomy and enable us to flourish. After mulling over various ideas, a grand plan begins to form in your head. You come to the group with your master plan, which you've dubbed "P," which lays out each member's responsibilities. You argue that, if each person follows P, he'll do the best that he

can, as an individual, to respect everyone's autonomy and to create the conditions for everyone to flourish.

After hearing your case, almost everyone is prepared to follow you. However, there is one dissenting voice. She argues that if everyone follows P, life in the supposedly utopian community will go quite poorly. She then presents a plan of her own, "P*," and explains that, collectively, the group would do significantly better in light of considerations of autonomy and flourishing if everyone followed P*. It is clear from people's responses that many members are being won over by this new plan, yet you listen to the presentation with a knowing smile on your face. When your rival is finished, and the spotlight turns back to you, you give the following response: "Of course, it would be far better if *all* of us followed P* instead of P. I never denied that. What I said was that *each* of us ought to follow P."

The other members of the commune are underwhelmed by this answer. They protest that your plan is *self-defeating*. By your own lights, if everyone follows P, they'll do much worse than they would have done by following P*.

Like the other members of this commune, many philosophers are skeptical about self-defeating normative theories. Across a wide range of debates in ethics, decision theory, political philosophy, and formal epistemology, philosophers have assumed that if a normative theory is self-defeating, then that fact alone shows that the theory is defective.[1] The burden of this paper is to show that this widely shared and consequential assumption is false. We have no reason to reject a theory simply on the grounds that it is self-defeating. Rather than focusing on some particular normative domain (such as morality or individual rationality), I will consider the phenomenon of self-defeat primarily at a more general level. In particular, I'll argue that a self-defeating theory can still meet key desiderata for a *domain-general* normative theory of what we ought to do, think, feel, etc. *all things considered*.

2 | SELF-DEFEAT

Let's begin with a concrete example in which a normative theory is self-defeating. Consider the following case, familiar from game theory:

Prisoner's Dilemma: Andy and Betty have been arrested for robbing the local bank and placed in separate holding cells. Each must decide whether to defect by ratting the other out or cooperate by remaining silent. If both cooperate, each gets one year of prison. If both defect, they get seven years each. If one defects and the other cooperates, the defector goes free and the cooperator gets ten years. Each prisoner prefers to spend as little time in jail as possible.

Consider *self-interest theory*, which says that an agent ought to act so as to make her life go as well as possible, and assume that each prisoner's life will go better, the less time he or she

[1] I'll introduce various examples from these literatures in section 2.

spends in jail. In this case, defecting *strictly dominates* cooperating. That is, each prisoner gets a better outcome by defecting, regardless of what the other prisoner does. If Betty defects, Andy gets seven years of jail time by defecting, instead of the ten he'd get by cooperating. If Betty cooperates, Andy will go free immediately by defecting instead of spending a year in jail. Betty faces the same prospects. So each prisoner ought to defect, according to our theory. Yet it's also true, on this theory, that it's much better for *each* prisoner for *both* to cooperate than to defect, given that each prisoner will spend seven years in jail if they defect, but only one year in jail if they cooperate.

Self-interest theory is *directly collectively self-defeating* in Parfit's sense that there are situations in which "it is *certain* that, if we all successfully follow T [the theory], we will thereby cause the T-given aims *of each* to be worse achieved than they would have been if none of us had successfully followed T" (1984, 55, emphasis in original).[2] A theory's T-given aims are its axiology, which gives us a ranking of the outcomes available to an agent or set of agents.[3] In the Prisoner's Dilemma, it is certain that if each prisoner follows our theory by defecting, his or her life will go worse than it would have gone if each of them had remained silent.

Most philosophical discussions of the Prisoner's Dilemma accept that the correct theory of individual self-interest is directly collectively self-defeating in this case. Notably, philosophers have been much less accepting of self-defeat in other contexts. In many philosophical debates, philosophers will offer a case that shows that some normative theory is self-defeating, and conclude solely on this basis that the theory should be rejected.

We see this style of argument in debates about the rationality of various attitudes, including *time biases* (that is, preferences for certain events over others based simply on when the events in question occur), *intransitive preferences* (preferring A to B, B to C, and C to A), and *imprecise preferences* (*not* preferring A to B or B to A, while also not being *indifferent* between them).[4]

As a fairly simple illustration, consider *present aim theory*, which directs each agent to satisfy his present preferences to the greatest extent possible. Note what this theory says in a case where an agent's preferences shift:

> **The Russian Nobleman**: You are a 20-year-old fervent left-winger. But you know that by middle age, you will become an equally fervent right-winger. You will receive

[2] A few notes of clarification. First, the definition here concerns *direct*, *collective* self-defeat. I'll have more to say about the significance of these qualifiers below. Second, let's assume that the agents in these cases know all of the relevant facts and can reason perfectly well about them, so that what they *subjectively* ought to do, given their relevant beliefs and/or evidence, coincides with what they *objectively* ought to do, given the facts about the situation (Or, if you prefer, assume that the agents' levels of confidence or credence in the relevant facts are sufficiently high for the subjective and objective oughts to align.). Third, by "normative theory," I mean to pick out both full-fledged normative theories such as act utilitarianism, that settle (or aspire to settle) questions about what to do in every choice situation, as well as much less comprehensive theories, which don't cover all situations. Fortunately, we often don't need to know very much about a normative theory to know whether it's self-defeating in certain cases.

[3] Parfit never defines T-given aims beyond saying that they are "what [normative theories] direct us to try to achieve" (1984, 3), and he says that they can include non-consequentialist considerations such as refraining from actions that are strictly prohibited and respecting others' rights. He also talks about T-given aims being *better* or *worse* achieved as a function of what the relevant agent does and the state of the world, which strongly suggests that T-given aims are a theory's axiology. Additionally, he's clear that a theory's aims might be better achieved without the agent trying to follow the theory, indicating that, despite the language quoted above, T-given aims are a theory's criteria for evaluating outcomes, rather than goals that the theory directs the agent to adopt for herself.

[4] See, e.g. Davidson et al. 1955 on intransitive preferences; Hedden 2015 (2.4) on imprecise preferences (Hedden's argument parallels an argument from Elga (2010) that theories of rationality that allow for imprecise *credences* can be self-defeating.); Dougherty 2015 and Sullivan 2018 on time biases.

> an inheritance of $100,000 at age 60. Right now, you have the option (call it Donate Early) of signing a binding contract which will require $50,000 to be donated to left-wing political causes. No matter whether you take this option, you will at age 60 have the option (call it Donate Late) of donating $50,000 to right-wing political causes (No greater donation is permitted under Tsarist campaign finance laws.). At age 20, your possible combinations of choices rank for you as follows: (1) Donate Early and Don't Donate Late; (2) Don't Donate Early or Late; (3) Donate Early and Late; (4) Don't Donate Early and Donate Late. At age 60, (1) and (4) are swapped, but you still prefer (2) to (3). You'd rather not donate at all than donate to both causes (since the effects of your donations would then cancel each other out).[5]

Your preferences in The Russian Nobleman place you in a Prisoner's Dilemma with yourself. At 20, you most prefer Donate Early, regardless of what you do at 60. At 60, you most prefer Donate Late, whatever you've done at 20. Yet at each time, you prefer that you never donate than that you donate to both sides. So present aim theory is what Parfit calls *individually directly self-defeating*, in that there are circumstances where "it is certain that, if someone successfully follows T, he will thereby cause his own T-given aims to be worse achieved than they would have been if he had not successfully followed T" (1984, 55).

Formal epistemologists have argued for various principles of rationality on the grounds that, without such principles, rationality is sometimes individually self-defeating. For example, Bayesians hold that you ought to conditionalize on your evidence. That is, upon learning some proposition E, you must set your level of confidence in every other proposition P to your level of confidence in P conditional on E. Lewis (1999) shows that if you fail to conditionalize, you can face a series of bets such that your best option at each choice point is to accept the bet, yet accepting all of the bets guarantees you a loss.[6,7]

In Prisoner's Dilemma and Russian Nobleman, it is *certain* that the agents' T-given aims will be worse achieved by following the relevant theory. That is, there is only one set of actions that constitutes following the theory (Defect, Defect; Donate Early and Late), and the theory's aims would have been better achieved if some different set of actions had been performed. A theory T is *possibly directly self-defeating* if and only if it is *possible* for the relevant agents to successfully follow T, even though their T-given aims would have been better achieved if they had acted in some other way (including following T in some other way).[8] This definition is broader than Parfit's definitions of direct individual and collective self-defeat, since it covers cases in which there is only one way to follow the theory (like Parfit's definitions), and cases in which there are multiple ways to follow it.

[5] This example comes from Hedden (2015, 424), and is a variation on a well-known case from Parfit (1984, 327). Hedden grants that rationality can be self-defeating (though he uses somewhat different language to express this point).

[6] Examples such as Lewis' are complicated, because one could simply deny—as various philosophers have—that Bayesianism and other theories of rational credence rank outcomes based on *practical* considerations, such as the avoidance of sure monetary losses.

[7] For many other such cases where various theories of rationality are individually directly self-defeating, see Hedden 2015. See also Dougherty 2015 and Sullivan 2018 for additional cases where time biased theories are self-defeating.

[8] The term *possibly directly self-defeating* is original to me. Parfit discusses theories that are self-defeating in this way when he presents his initial definition of a directly collectively self-defeating theory (1984, 53), and when he considers the possibility of agents following consequentialism in a suboptimal way in coordination cases (72). He never offers a definition of this kind of self-defeat, though my definition fits well with what Parfit says about coordination problems.

The (mere) possibility of self-defeat arises when a theory generates coordination problems, where what someone ought to do depends on what is done at other choice points. Consider the following two cases:

> **Professor Procrastinate**: Professor Procrastinate is asked to review a graduate student's paper, soon to be given as a job talk. He is the best person to review it and has ample time to do so. If he gets his comments to the student on time, she'll give an amazing talk and likely receive a job offer. Unfortunately, he's Professor Procrastinate. He knows that if he accepts the request, he won't review the paper in time for the student to respond to his comments, and the talk will go terribly. If he declines the request, she'll ask Professor Punctual, who will offer her timely yet mediocre comments, and she'll then go on to give a mediocre talk.[9]

> **Slice and Patch**: Unless a patient's tumor is removed very soon, she'll die (though not painfully). The only way to save her is for the two available doctors, Slice and Patch, to come down to the hospital immediately and perform surgery. Neither doctor can save the patient on his own, and it would even be cruel for only one of them to show up, as this would get the patient's hopes up and make her death psychologically agonizing. Unfortunately, Slice knows that Patch won't show, even if Slice does. Patch is going to stay home to tend to his child, who's suffering from a bad (though in no way life-threatening) case of the flu. Patch also knows that Slice won't show, regardless of whether Patch does, because Slice is going to stay home to tend to his child, who (coincidentally) is also suffering from a non-life-threatening case of the flu.[10]

Suppose we accept *actualism* and think that, morally, the outcome where the student gives an amazing talk is better than the one where she gives a mediocre talk, yet whether Professor Procrastinate should accept or decline the invitation depends on whether he's going to write the review.[11] That is, Procrastinate can follow our theory either by declining the request, or by accepting and then writing. Our theory is only possibly self-defeating, then, because Procrastinate can follow it in a way that best realizes the theory's aims. Still, he can follow our theory even if he's completely unwilling to achieve its aims to the greatest extent possible. All he has to do is decline the request.

Similarly, in Slice and Patch, we might think that it's far better, morally, for the patient to be saved by the doctors than for each doctor to tend to his own child, but also that whether each doctor should come in for surgery or tend to his child depends on what the other doctor does. The doctors can follow our theory either by doing their respective parts in the surgery or by tending to their respective children. Moreover, even if each doctor is obstinate in his unwillingness to satisfy the theory's aims the best that he can, Slice and Patch will each nonetheless follow the theory by tending to their children.

[9] The original version of this case comes from Goldman (1978). There are many variations in the literature. My version is taken (with minor modifications) from Timmerman and Cohen (2020).

[10] This case is introduced by Estlund (2017). See also his 2019 (especially ch. 11) for further discussion. My version borrows some details from Portmore's discussion (2018; 2019, ch. 5).

[11] More abstractly, *actualism* is the view that an agent is obligated to φ if and only if what would happen if she φed is better than what would happen if she didn't φ. *Possibilism* is the view that whether an agent ought to φ is a matter of whether φ is a member of the best set of acts that she can perform.

Parfit says that the possibility of self-defeat is less objectionable than the certainty of it (54). However, many other philosophers have rejected normative theories simply on the grounds that they can be self-defeating in coordination problems.[12] I'm primarily interested in whether we should reject a normative theory on the grounds that it is self-defeating in *any* of these senses. For most of the remainder of this paper, then, when I talk about a theory being "self-defeating," I mean that it is *possibly directly self-defeating*, since this definition is capacious enough to cover cases where self-defeat is certain or only possible, at either the individual or collective level.

Taking this approach will, I believe, prove fruitful. If successful, my attempt to defend self-defeating theories will exonerate theories that are possibly, certainly, individually, and/or collectively directly self-defeating. That is, if I'm right, the fact that a normative theory is directly self-defeating, as such, gives us no reason to reject it.

In my examples in this section, I've made reference to theories of specific normative domains, such as morality and self-interest. This is in keeping with the existing literature on self-defeat. However, my primary interest is in self-defeat as a domain-general phenomenon. I'm interested in whether self-defeat is a problem for a normative theory as such, i.e. as an internally consistent theory that can guide action and serve as a basis for criticism. I want to show that the fact that a domain-general theory is self-defeating, by itself, is no objection to the theory.[13] In section VI, I'll address Parfit's argument that the best theory of morality cannot be directly collectively self-defeating.

Before getting into the arguments, I should make one note about the scope of this paper. Following Parfit, I've been talking in terms of *directly* self-defeating theories. A normative theory is *indirectly self-defeating* if and only if it's the case that were the relevant agents to *try* to follow T, their T-given aims would be worse achieved than if they hadn't tried to follow it.[14]

The paradox of egoism is one example of indirect self-defeat. Consider again self-interest theory, which says that you should do whatever makes your life go best. Trying (in each action) to follow this theory will lead you to have a life that is less good for you than one that you could have had if you hadn't always aimed to do what was best for you. This is because, for instance, genuine friendships and loving relationships are among the greatest contributors to our well-being, but having purely self-interested motivations precludes us from having loving relationships and friendships.

Although indirect self-defeat raises a host of interesting questions, my focus is on theories that are directly self-defeating, for three reasons. First, indirect self-defeat can arise simply from limitations in our information and in our ability to process information. For example, one reason why you might produce better consequences by following various rules of thumb instead of always doing what you think will produce the best consequences is that your beliefs about what will produce the best consequences get the wrong answer more often than the rules of thumb.

[12] For examples from the literature on self-interest, see, e.g. McClennen (1990, ch. 8-9) and Gauthier (1994). For examples from moral philosophy, see, e.g. Regan (1980), Zimmerman (1996, ch. 9), Tuck (2008), Pinkert (2015), Nefsky (2017; 2019), Portmore (2018; 2019 ch. 5), Fanciullo (2021), and Soon (2021).

[13] Existing defenses of self-defeating theories all focus on theories of morality (or of rationality in the case of Christensen 1991 and Hedden 2015). See, e.g. Feldman 1980; Jackson 1987; Kierland 2006; Preston-Roedder 2014; Budolfson ms.

[14] My definition is taken from Parfit (1984, 5, 27), with slight modifications so that it covers both individual and collective indirect self-defeat.

In contrast, theories can be directly self-defeating even when the agents have all of the relevant information and are ideal reasoners.

Second, the typical response to the fact that a theory is indirectly self-defeating is to distinguish between treating a theory as a *standard of rightness* and as a *decision procedure*. A number of philosophers think that certain normative theories, such as consequentialism, are defensible as a standard of rightness but not as a decision procedure. They think, moreover, that a standard of rightness should play a much more indirect role in deliberation.[15] This type of response may address the problem of indirect self-defeat, but it does not carry over to direct self-defeat. The problem of direct self-defeat is that agents can successfully follow a theory's standard of rightness and thereby act in a way that's worse—by the theory's own standards—than if they hadn't followed its standard of rightness. In other words, direct self-defeat looks like a problem that's internal to a theory's standard of rightness, unlike the problem of indirect self-defeat.

Finally, as Wiland (2007) points out, there are a variety of ways in which a normative theory can be indirectly self-defeating (depending on how directly the theory requires you to try to follow it). For example, you might bring about an outcome that's bad by the theory's lights if you incessantly focus on following the theory. Alternatively, you might cause a bad outcome simply by accepting the truth of the theory. I'm wary of extending my argument to cases of indirect self-defeat because it's unclear to me that we should say the same thing about each kind of indirect self-defeat.[16]

3 | THE FIRST ARGUMENT AGAINST SELF-DEFEATING THEORIES: SUCH THEORIES ARE INCONSISTENT

The first general argument against self-defeating theories to consider is that these theories are inconsistent. Parfit gestures at this when he says that self-defeating theories "condemn themselves" (2011, p. 306). We can spell out the argument as follows:

> P1 A normative theory is inconsistent if there are situations where no matter what the relevant agents do, they will have done something that, according to the theory, they ought not to have done.
>
> P2: If a normative theory is self-defeating, then in some cases, no matter what the relevant agents do, they will have done something that, according to the theory, they ought not to have done.
>
> C: Self-defeating theories are inconsistent.[17]

[15] For instance, Railton's *sophisticated consequentialist* does not normally attend to whether his actions produce the best consequences, but he would not act for a given non-consequentialist reason if so acting was incompatible with leading a life that conformed to the consequentialist standard of right action (1984, 152).

[16] Perhaps, as Wiland argues (2007, sec. III), the fact that a theory is indirectly self-defeating in certain ways does, by itself, constitute a reason to reject the theory.

[17] My presentation of this argument is essentially lifted from Hedden 2015 (433).

P1 is easy to motivate. A normative theory is a theory of what to do. It tells us how to choose among our options. If there are situations where whatever agents do, they'll do something that they ought not to have done, by the theory's lights, then it looks like the theory isn't giving them a consistent answer to the question of what to do.[18]

What about P2? In all of our cases where some theory is self-defeating, we've assumed that the agent or agents ought to give *each* of a set of responses that realizes the aims of the theory worse than some alternative set of responses. The crucial assumption underlying P2, then, is that in these cases, the relevant agents ought not to give this entire *set* of responses, given that they could better realize the theory's aims through a different set of responses. Andy ought to defect and Betty ought to defect, but Andy and Betty ought not {defect, defect}. Similarly, the Russian Nobleman ought to Donate Early and he ought to Donate Late, but he ought not {Donate Early, Donate Late}. And so on for our other cases. If we grant this assumption, then we get the result that in circumstances where a normative theory is self-defeating, the agents cannot avoid doing something that they ought not to have done.

One way to resist this argument, of course, is to deny the assumption supporting P2, that the agents ought not to perform the set of actions that worse realizes the theory's aims. I think that there are fairly plausible grounds for rejecting this assumption.[19] However, if we take it on board, then we should reject P1. It's perfectly consistent for a theory to say that each individual response in some set of responses ought to be taken, given the relevant alternatives to each, and that the entire set of responses ought not to be taken, given *its* relevant alternatives. The option sets on which these verdicts are based differ, so there's no inconsistency.

When we assess each individual response, we take what happens at the other choice points as part of the circumstances. That is, the option set only includes options available at that choice point. In contrast, when we adjudicate between entire sets of responses, we're trying to settle what is to be done across all of the choice points, so the option set now includes each combination of individual responses.[20]

This is fairly abstract, so let's illustrate the point by returning to the cases from section II. In the Prisoner's Dilemma, self-interest theory says that each prisoner ought to defect, given that the other prisoner defects, and given that the other prisoner cooperates. This is consistent with the verdict that both prisoners ought to cooperate, if we interpret this second judgment as being about what the two prisoners should do, given the option set {universal cooperation, universal defection}. In other words, we can grant that the prisoners ought to cooperate rather than defect, even

[18] Some philosophers who believe in the possibility of moral dilemmas yet still think that the best theory of morality will be consistent have argued that a normative theory can be consistent even if there are some circumstances in which it cannot be fully satisfied (see, e.g. Marcus 1980).

[19] Hedden 2015 offers one such argument. He argues that an agent's options at some time are simply the *decisions* that she is able to make at that time (441). He contends that in all diachronic Prisoner's Dilemmas, there is no point at which the agent ought to decide to perform the set of responses that avoids the bad outcome. At each choice point, she ought to decide to give her best response at that choice point (sec. 5). Hedden's view of options is contentious and quite a departure from the commonsense assumption that non-mental actions can be options. Additionally, Hedden is clear that his account is only meant as an account of what our options are as a matter of *subjective rationality*, so it's not obvious whether his response generalizes to, e.g. self-defeating moral theories or self-defeating theories of all things considered normativity.

[20] This way of relativizing normative judgments to option sets is defended by Frank Jackson in his discussion of individual and group morality (1987) and of the actualism-possibilism debate (see Jackson and Pargetter 1986; Jackson 2014). Interestingly, Parfit also defends this position in several of his writings when discussing consequentialism's implications in coordination problems. For example, in an unpublished manuscript (1988), he writes, "When I ask what *I* should do, what *you* do is *part* of the circumstances…When I ask what *we* should do, what you do is *not* part of the circumstances" (7). See also 1984, 73.

though, if we just ask what Andy ought to do and take Betty's response as given, we get the answer that Andy ought to choose defection over cooperation in a world in which Betty cooperates, and in a world in which Betty defects (*mutatis mutandis* for Betty).

The same analysis applies in the Russian Nobleman. Given your preferences at age 20, present aim theory tells you to Donate Early, no matter what you do at 60. As someone with the preferences of a 60-year-old right-winger, it tells you to Donate Late, regardless of what you've done at 20. Nevertheless, it's also true, given either set of preferences, that as between {Donate Early, Donate Late} and {Don't Donate Early, Don't Donate Late}, you ought to do the latter. This judgment is consistent with the first two judgments, because the option sets are different.

In Professor Procrastinate, actualism says that Professor Procrastinate should decline the invitation, given that he isn't going to review the paper. This is compatible with saying that he ought to accept the invitation and then review the student's work, as opposed to declining and not reviewing it. This second judgment treats accepting and then writing as one of Procrastinate's options, whereas the first judgment does not. The first verdict treats the fact that Procrastinate won't follow through as simply part of the circumstances in which he chooses whether to accept or decline.

Similarly, in Slice and Patch, it's tempting to say that each doctor ought to stay home and tend to his child, given that the other doctor is going to do so. In arriving at this conclusion, we imagine each doctor in a world where the other doctor stays home, and ask whether the doctor whose future conduct is under scrutiny ought to stay home or show up for surgery in that world. In contrast, when we conclude that the doctors ought to perform surgery, we're assuming a world in which it isn't already set in stone what either doctor is going to do.

In all of these cases of self-defeat, we can take what we might call the Act Perspective or the Pattern Perspective.[21] The Act Perspective tells an agent what to do at a particular choice point, given all of her information about what may happen at the other relevant choice points. In contrast, the Pattern Perspective points out that some alternative set of responses across all of the choice points is available, and says that, as between this set of responses and the set of responses that we will give if we follow the Act Perspective, we ought to go with the former. There is no inconsistency when the verdicts of these two perspectives diverge, since their verdicts are based on different option sets.

4 | THE SECOND ARGUMENT AGAINST SELF-DEFEATING THEORIES: SUCH THEORIES FAIL TO BE ACTION-GUIDING

Imagine that you're Andy in the Prisoner's Dilemma. You're told that, with regards to your own self-interest, you ought to defect, even though you and Betty ought to cooperate rather than defect. Suppose that you accept everything that I argued in the previous section, and so you recognize that the Act Perspective's verdict takes Betty's behavior as given, whereas the Pattern Perspective's verdict does not, so the two are consistent. Nevertheless, you can only act on one of them. So even if self-defeating theories are consistent, they seem to fail to be action-guiding.

One way to try to evade this objection would be to posit some kind of collective agent (or diachronic agent) and to say that, in cases where a theory is self-defeating, the individual agents (or time-slices of agents) ought to follow the Act Perspective, whereas the collective agent ought to follow the Pattern Perspective. However, the only way for the group agent to act in the cases that

[21] I take the distinction between "acts" and "patterns" from Woodard 2008. In his discussion of diachronic cases, Wu (2022) draws a similar distinction between the "Immediate Perspective" and the "Planning Perspective" (sec. III.A.).

we've considered is through the individual agents. The only way, for instance, that the prisoners collectively can cooperate is by each prisoner cooperating. So if the prisoners collectively ought to cooperate through individual acts of cooperation, *and* these individual acts are subject to the Act Perspective, then we've made no progress in determining which perspective should carry the day.

The only way out of this problem is to show that, in any given situation, only one perspective is action-guiding.[22] I think that the Act Perspective is *always* action-guiding. I'll give a general rationale for this position, but first, it will be helpful to get on the table some cases where it seems undeniable that you should act in accordance with the Act Perspective.

4.1 | Cases

Consider the following interpersonal case:

> **Dictatorship**: 100,000 of us (all able-bodied adults) are living under a dictatorship. Each one of us is made to wear a bracelet that allows the government to track our movements and shock us if we are unruly. Shackled with these bracelets, it is impossible for anyone to revolt. However, we all know that if the bracelets were deactivated, we would be able to overthrow the regime if *everyone* revolted. One Friday afternoon, our bracelets suddenly deactivate. A broadcast is sent out, enjoining everyone to remain calm and to carry on with their assigned tasks, and assuring us that the problem will be fixed in a matter of minutes. The broadcast ends on an ominous note, warning that any citizen who removes her bracelet will be subjected to years of agonizing torture and eventually executed. There is no time for us to communicate with one another. If we are to overthrow the government, each of us must act now, and hope that our fellow citizens do their part.

Each citizen has two options: remove her bracelet or keep it on. The (morally and self-interestedly) best thing for all of us to do is to remove our bracelets so that we can get the revolt underway. As between everyone removing their bracelets and everyone keeping them on, it's clear that we should all remove our bracelets. But does this Pattern Perspective verdict have any bearing on whether each individual should remove her bracelet?

Intuitively, no. Each of us should act based on what others are going to do. Given that if *anyone* keeps her bracelet on, then every other citizen should do the same, and the terrible consequences of removing your bracelet if not everyone makes this choice, it's extremely plausible that each citizen knows that someone is not going to remove her bracelet. Even if each of us resolved to do so, somebody would lose her nerve, or not remove her bracelet because she expects one of her fellow citizens to lose her nerve, etc. So even though we can all remove our bracelets and this is far and away the best course of action, it seems that each individual citizen should give her best response according to the Act Perspective. That is, she should keep her bracelet on, given what she knows about what others are going to do.

Let's turn now to a diachronic case involving a single agent:

[22] Lazar and Lee-Stronach (2019) and Wu (2022) also defend the view that only one perspective can be action-guiding in a given situation (the former refer to "acts" and "campaigns"), though they each think that which perspective is action-guiding depends on the circumstances.

> **Summer Chores**: Your parents promise to buy you a car if you complete an extremely long list of 1,000 chores for the summer. You recall that when your brother completed all but one of his chores, your strict parents didn't buy him a car. You somewhat prefer doing all of the work and getting the car to no chores and no car. However, you loathe chores (you're a selfish brat and don't care at all about helping your parents). Except for the last chore, you always prefer doing fewer chores to more. You know that you have the requisite skills and time to complete your chores, but throughout the summer, you'll be faced with many temptations, and it will take considerable effort to be disciplined enough to finish everything.

You're now faced with the decision of whether to start working on your chores or go tubing out on the lake with friends. What should you do?

You should go tubing.[23] Given the number of chores you have to complete to get the car, you know that you're very likely to fail in this endeavor. First, there is the possibility that you will simply slip up at some point. Second, there is the possibility that you will anticipate that you will later slip up, and decide to stop based on this expectation.[24] Moreover, your preference for getting the car isn't *so* strong (given how much you despise chores) as to justify making the effort based on the slim chance that you succeed in getting through everything.

Yet it remains the case that you prefer doing your chores and getting the car to not doing them and going into your senior year carless. So the Pattern Perspective tells you to perform the sequence of actions in which you complete your chores. If we think, then, that you shouldn't even start them, that judgment reflects the Act Perspective.

4.2 | Why Only the Act Perspective is Action-Guiding

What is the best explanation for why the Act Perspective is action-guiding in cases like Dictatorship and Summer Chores? As I explained in section III, the reason that the Act and Pattern Perspectives diverge is that the former takes more things as given. In particular, the Act Perspective takes as given anything that may or will be done at other choice points, whereas the Pattern Perspective treats certain combinations of actions across all choice points as options. To decide which perspective is action-guiding, then, we need to decide whether information about what other agents can be expected to do *should* be taken as given.

The best reason for taking this information as given, which applies to all of our cases, is that it's *known* by each of the agents at the various choice points (or each agent has a sufficiently high justified credence in the relevant propositions). In other words, the Act Perspective takes it as given that certain things will happen (or have a certain probability of happening) at other choice points, when this information is available to the agent, whereas the Pattern Perspective ignores these givens. Examples such as Dictatorship and Summer Chores help us see that this information should be taken into account, because it's so clear in these cases that any agent who fails to do so is courting disaster.[25]

[23] At least, you should go tubing as a matter of self-interest. Perhaps you have some filial obligations that require you to start your chores, even if you know that you won't finish them.

[24] Or you know that you will later anticipate that you will give in, and so on.

[25] As I'll explain in section V, I also think that these cases are clearer because there isn't a worry about whether taking this information as given for purposes of action-guidance lets the agents off the hook too easily.

Dictatorship and Summer Chores are not Prisoner's Dilemmas. In the former cases, there are uniquely best patterns available, and if each agent had sufficient reason to expect everyone else to play their roles in the patterns, then it would be the case that each agent at each choice point ought to act so that everyone gives the best set of responses. For this reason, it's worth explaining how the argument just given applies to Prisoner's Dilemmas, where it's never the case that the Act Perspective's verdicts align with those of the Pattern Perspective.

Let's focus on the original Prisoner's Dilemma. As a matter of self-interest, we assumed that the different patterns of action that the prisoners can perform rank (for Andy) as follows (for Betty, 1 and 4 are switched):

1. Andy defects, Betty cooperates
2. Andy cooperates, Betty cooperates
3. Andy defects, Betty defects
4. Andy cooperates, Betty defects

Pattern 2 is better than pattern 3 for both Andy and Betty. Why, then, should they act in accordance with the Act Perspective and settle for 3, when they could have reaped the benefits of 2 by following the Pattern Perspective? The problem with following the Pattern Perspective is that it ignores information about the circumstances in which Andy and Betty choose between cooperation and defection. Andy knows that, when he chooses whether to cooperate or defect, he will be in one of two possible situations, based on what Betty does. Either Betty will cooperate, or she'll defect. The Act Perspective tells Andy what to do in each of these scenarios, given what is best for him in each. If she cooperates, then his choice is between 1 and 2, and he should choose 1 by defecting. If she defects, then his choice is between 3 and 4 and, again, he should choose 3 by defecting. The Act Perspective's recommendation reflects the fact—which both perspectives can agree on—that 1 is better for Andy than 2, and 3 is better than 4. The Pattern Perspective, on the other hand, ignores this feature of Andy's situation. It treats each agent as if he or she could simply choose 2 over 3. Given that the prisoners act independently, that assumption is untenable.[26]

Think of the two perspectives as advisors. They have exactly the same values, yet they offer you conflicting advice. As you discover, the reason for this divergence is that one of them takes into account certain information (available to the advisers and to you) that the other adviser ignores. Clearly, you should listen to the first advisor, even if everything that the second advisor says is correct, as far as it goes.

Self-defeating theories, then, are action-guiding, because only the Act Perspective is action-guiding. If a theory's axiology is such that the Act and Pattern Perspectives can conflict, this feature poses no threat to its suitability for guiding action.

4.3 | Spectrum Puzzles and the Act Perspective

Having argued that only the Act Perspective is action guiding, I want to close out this section by considering a class of cases that has led many philosophers to think that, in fact, the Pattern

[26] This assumption might be plausible if the prisoners were choosing together. Neither one of them would plausibly choose to cooperate if the other defects, so universal cooperation and universal defection might effectively be the only options in such a scenario.

Perspective is sometimes action-guiding. In these cases, following the Act Perspective seems to lead to some pretty painful and repugnant places.[27] Consider:

> **Puzzle of the Self-Torturer**: The self-torturer is hooked up to a medical device that administers an electrical current into the body. The machine has 1,001 settings: 0 (off) up to 1,000. Each week, doctors give the self-torturer the following options: stay at his current setting or go up one setting. Every time he goes up, he receives $10,000, and the increase in shock is so minor that he always prefers to go up than to stay put. However, at setting 1,000, the pain of the shock is so strong that he'd gladly give up all of his money to have the machine turned off.[28]

For any two adjacent settings n and n +1, the self-torturer prefers n + 1 to n, but he also prefers setting 0 to setting 1,000. His preferences are *intransitive*.

Imagine that the self-torturer follows the Act Perspective and chooses his preferred option each week, ultimately ending up in a worse outcome (the machine is set to 1,000) than if he had simply stayed at 0. Most philosophers who have discussed this example assume that it cannot be the case that the self-torturer should go all the way up to setting 1,000, given that this pattern of action realizes his aims worse than many alternative patterns available to him.[29] One common approach to this puzzle is to argue that the self-torturer's choices should be guided by a rational *plan* that he adopts, or that it would be rational for him to adopt.[30] Since the plan tells him which set of choices to make across all choice points, the self-torturer's choices are guided by the Pattern Perspective, rather than the Act Perspective, on this approach.

It's easy to think that, were he to adopt a plan to follow a certain pattern of action, the self-torturer wouldn't end up going to 1,000, because he would surely choose (e.g.) the pattern in which he stays at 0 over the pattern in which he ends up at 1,000. White (2015), for instance, is quite explicit that whether a given pattern is rational for the self-torturer to follow depends on whether he prefers the outcome of following this pattern to the outcome of never turning the machine on. He says, "If what [the self-torturer] is doing is just part of proceeding from 0 to, say, 400…then if he is rational, he will only move to 400 if there is good reason for him to prefer the combination of money and discomfort at 400 to his initial impoverished, though physically comfortable, state" (604).

However, if we evaluate the self-torturer based on the pattern that he follows, the comparison between 0 and 400 is not the only relevant comparison. We should also ask how stopping at

[27] I'd like to thank an anonymous referee for raising this objection.

[28] This case originally comes from Quinn 1990. Other common examples of spectrum puzzles include Parfit's Drops of Water case (1984, 76) and the "lawn-crossing" problem (see, e.g. Rabinowicz 1989, sec. 10). Parfit's argument for the Repugnant Conclusion (ch. 17) in population ethics can also be seen as a spectrum puzzle (though it is usually treated primarily as a puzzle in axiology, rather than as a puzzle for how we ought to choose). Additionally, a number of philosophers have argued that climate change is a high stakes spectrum puzzle. See, e.g. Andreou 2006.

[29] For example, White says that one constraint on an adequate solution to the puzzle is that it "must explain why going all the way to 1000 is irrational in all cases where the self-torturer has the relevant preferences [i.e. he prefers the higher setting for any two adjacent settings and prefers 0 to 1,000] and is fully informed about the relevant facts" (2015, 588).

[30] For developments of this approach that rely on the self-torturer actually forming a plan, see, e.g. Quinn 1990; Bratman 1999; Andreou 2006. For versions that don't require the self-torturer to form a plan, see Carlson 1996; White 2015.

400 compares to stopping at 399, for instance.[31] Given that the self-torturer's preferences are intransitive, *any* pattern that he chooses will look bad in light of some of his preferences. For any pattern in which he stops before 1,000, there is at least one alternative (in which he goes up another setting) that he prefers. Additionally, of course, if he follows the pattern that results in him stopping at 1,000, then he'll wish that he had stayed at 0. So while following the Act Perspective may guarantee that the self-torturer ends up in a bad outcome, the Pattern Perspective fails to offer him coherent guidance about which pattern to instantiate. Moreover, the only way that it can generate coherent advice is by *ignoring* certain pieces of information about how the various possibilities compare to one another (e.g. by treating the fact that 0 is preferable to 400 as a decisive reason to stop short of 400, but ignoring the fact that 400 is still preferable to many other settings (399, 398, and so on).

However the self-torturer approaches his situation, he'll fail to get what he wants because there is no possible world in which he gets what he wants.

5 | DOES THE ACT PERSPECTIVE INVITE IRRESPONSIBILITY?

You might object that there's something dubious about the particular *kind* of additional information that the Act Perspective takes into account. Namely, facts about what you or other agents may or will freely choose to do.

Uneasiness about taking facts about agents' choices as given pervades work on both diachronic agency and collective action. Nefsky writes that "in contexts of practical deliberation…we think of agents (both ourselves and others) as typically being able to choose between several different courses of action" (2017, 2762). She concludes from this observation that our conception of what's possible for purposes of deliberation "cannot be one that, in general, holds fixed what agents will choose to do in the future" (id). Moran (2002) argues, in the context of diachronic agency, that to predict that you'll act in a certain way (based, e.g. on psychological or behavioral evidence), and then use that prediction to justify your present choice, is an evasion of responsibility.[32] Questions of the form "how *will* I act" should be answered by resolving the question, "how *am* I to act?"[33]

There are two concerns expressed in these quotes. One is that we shouldn't settle deliberation about what to do on the basis of information about what we'll in fact do. The other is that we shouldn't appeal to facts about how we'll choose to avoid accountability for those choices. I will now argue that even though an agent's choice in a given situation should take into account choices

[31] Since the self-torturer begins at setting 0, it's natural to treat this as the baseline against which to compare each of the other settings that the self-torturer could choose as his stopping point. However, the fact that he starts at 0 intuitively shouldn't have this kind of significance. Suppose that the self-torturer began at setting 500, and could choose each week to go up one setting, until he reached 1,000, at which point he could choose to go to 0, and then to keep going up to setting 499. Given that he has all of the same options in this version of the case as he had in the original, and that his preferences are exactly the same, what's rational for him to do should be the same in each situation. However, if we compare each of the other settings solely against the self-torturer's starting setting, we'll likely end up giving him different recommendations about where to stop in each of these versions. For example, perhaps he should stop at 300 instead of remaining at 0, but he should stop at 700 instead of staying at 500.

[32] Moran grants, however, that it can make sense to act in light of your predictions of your own future behavior if you doubt that you'll stick to your resolutions (94).

[33] Marušić (2015) also develops a view along these lines, though he is hard to pin down, since he thinks that you can remain in the deliberative stance but still indirectly take into account the risk of failure by regarding the *difficulty* of performing some sequence of actions as a reason not to perform it (ch. 6.1).

that she or other agents make at other choice points, she may nevertheless bear responsibility for some of those choices. This point is easiest to show in the case of Professor Procrastinate, so I'll start there, before turning to the other cases.

5.1 | Professor Procrastinate

Suppose that Procrastinate knows that if he simply started reading the student's paper, he'd quickly become engrossed in his work and it would take little effort to remain engaged long enough to produce helpful, insightful comments. However, he also has excellent reason to believe that he won't even make this minimal effort. Whenever an opportunity to review the paper presents itself, he'll instead put on an old James Bond movie. If this is Procrastinate's situation, do we really want to say that he does exactly what he ought, all things considered, to do if he declines the review?

The trouble with answering "yes" to this question—as the Act Perspective does—is that it seems to commit us to the view that Procrastinate is in no way blameworthy for his course of action. However, we need not take on this further commitment. We can accept that the ought of the Act Perspective is the all things considered, action-guiding ought, while also holding that the *some things* considered ought of the Pattern Perspective has a role in the practice of accountability.

In the last section, we imagined the Act and Pattern Perspectives as advisors. I argued that the best advisor will take into account all of the relevant information available to her. Now, let's think of these two perspectives as judges deciding whether Procrastinate is to be sanctioned in some way for declining to help the student. Procrastinate argues that he's not at fault. He did the best that he could in declining, given that he was never going to even try to review the paper. The first judge, personifying the Act Perspective, accepts this defense and rules in Procrastinate's favor. The second judge, embodying the Pattern Perspective, rejects it and rules against Procrastinate. Given that Procrastinate could have easily reviewed the student's work, the fact that he knew that he wasn't going to does not absolve him of responsibility.

Although the Act Perspective is a better advisor, here I think that *each* perspective gets something right in its capacity as judge. Given that Procrastinate isn't going to review the paper, he's accountable for taking this fact into consideration in deciding whether to accept the invitation. Certain people, such as the student, would be entitled to blame him if he accepted the invitation knowing that he wasn't going put to in any effort towards writing it. Nevertheless, it's also true that he is accountable for reviewing the paper, given that he easily can. If he fails to do so, then he's blameworthy for this failure.[34]

How can both perspectives get something right about what Procrastinate is on the hook for, even though only one perspective is correct about what he ought, all things considered, to do? The reason that only one perspective can be right in the advice context is that it's only possible for Procrastinate to take one perspective's advice. If he declines, he follows the Act Perspective

[34] Given the tight connection many philosophers posit between blame and *motives*, it's worth noting that, although we can blame Professor Procrastinate for his failure to conform to a certain pattern, we need not think that his motives to act must in some way reflect the Pattern Perspective. He can conform to the requisite pattern even if he isn't motivated by pattern-based considerations as such, and is motivated solely by act-based considerations. Provided that he actually writes the review, he can accept the request simply for the reason that accepting is better than declining, given that he's going to write, and he can then go on to write on the grounds that writing is better than not writing, given that he has accepted. I'd like to thank an anonymous referee for pressing me for clarity on this point.

but cannot follow the Pattern Perspective. If he accepts, he takes the first step in following the Pattern Perspective but can no longer follow the Act Perspective. In contrast, there is nothing that prevents someone, such as the student, from blaming Procrastinate if he accepts the request and blaming him if he doesn't review the paper.[35]

Even if it's *possible* to blame Procrastinate for each of these things, you might still protest that it's unfair to put him in a bind where, no matter what he does, he'll be blameworthy for something. The thing to say here is that holding Procrastinate accountable in this way is fair, given that he puts himself in this bind. The only reason that he cannot avoid blame is that he's going to watch Bond films when he could easily write the review instead.

This response is analogous to Aquinas' resolution of moral dilemmas.[36] Aquinas thought that there are situations where whatever you do, you'll do something morally wrong. However, he believed that all dilemmas are *secundum quid* dilemmas, that is, dilemmas that arise from some prior wrongdoing on the agent's part. Suppose that I find myself in a situation where I am unable *now* to avoid doing something wrong, because I have made two promises and cannot fulfill both. If I could have foreseen that I wouldn't be able to keep both promises when I made them, then it's plausible that I ought to break the lesser promise and that I'm blameworthy for doing so, given that I shouldn't have gotten myself into this situation. The case of Professor Procrastinate shows that this kind of moral dilemma can also arise from some *counterfactual* morally discreditable action that the agent would perform.

5.2 | Slice and Patch, Summer Chores, and Dictatorship

Can we hold agents accountable based on the Pattern Perspective in our other cases? First, let's consider the cases that share, with Professor Procrastinate, the feature that none of the agents has an option that's strictly dominant at any choice points. What each agent should do depends on what's done at the other choice points.

Slice and Patch differs from Professor Procrastinate in that the case for (say) Slice not showing up for surgery isn't based on what Slice himself is going to do, but on what another agent, Patch, is going to do. However, this detail doesn't significantly change the analysis. Slice ought, all things considered, to stay home, given that Patch is going to stay home. Moreover, Slice is accountable for staying home, given that he has this information about Patch. Slice's child, for instance, could plausibly blame him if he doesn't stay home. At the same time, Slice is going to stay home *regardless* of whether Patch comes in. So we should also blame Slice for his failure to do his part in saving the patient. Slice cannot avoid accountability, but as with Professor Procrastinate, he's in this bind through his own (counterfactual) doing. (Analogous points apply to Patch.).

Summer Chores is similar to the case of Professor Procrastinate in that there is only a single agent acting across time. Yet it's much more difficult for you to complete your chores than for

[35] Timmerman and Cohen (2016) present two hybrid actualist-possibilist accounts of moral obligation that are similar to the position I defend here. On their first model, Professor Procrastinate faces conflicting obligations, but the conflict is due entirely to his past failures to satisfy his moral obligations (sec. 5). On their second model, his only obligation is to accept the request and review the paper, but this obligation isn't action-guiding (sec. 6). Moreover, they suggest that even if an obligation isn't action-guiding, an agent's failure to meet it may still warrant blame (683). Put in the language of obligation, my view is that Procrastinate faces conflicting obligations when he is put in the position of accepting or declining the request, only one of which (the obligation to decline) is action-guiding, and the conflict is due to his counterfactual failure to satisfy one of the obligations.

[36] Or, at least, Aquinas as interpreted by Dougherty (2011, 138-39).

Procrastinate to conduct his review. Unlike Procrastinate, who can easily complete his task in a single afternoon if he tries, you must regularly decide to work on your chores (and follow through on your decisions) if you are to complete all 1,000 of them. You aren't blameworthy for the fact that you would slip up at *some* point in this long process, even though it's within your power, at each point, not to slip up.[37]

Dictatorship is a case where it's even more clear that no one is to blame for the fact that we fail to remove our bracelets and rise up in revolt. Again, it's not as though removing one's bracelet is difficult. However, given that all 100,000 of us must remove them for the revolt to be successful, and the dire consequences of removing yours if not everyone else does likewise, it seems cruel to blame any of us for the fact that we cannot count on everyone to remove his bracelet. Moreover, it is to each of our credit if we keep our bracelets on so as to avoid being pointlessly tortured and executed, and simply tragic that we fail to seize the opportunity to revolt.

I suspect that one reason why it's so intuitive to think that only the Act Perspective is action-guiding in Summer Chores and Dictatorship is that we're not worried about letting the agents off the hook in these cases. We already accept that they don't bear responsibility for failing to act in accordance with the Pattern Perspective. We have no trouble, then, in thinking that it's also inadvisable for the agents to follow the Pattern Perspective.

5.3 | Prisoner's Dilemma and The Russian Nobleman

In one-off Prisoner's Dilemmas, there is no plausible sense in which the agents are criticizable for the fact that their actions are self-defeating, as far as self-interest theory is concerned. Professor Procrastinate is blameworthy for not performing his *best* available sequence of actions when he easily could have. Andy and Betty face a different situation. For Andy, the best set of responses is for him to defect and Betty to cooperate. However, he's not in a position to choose this set of responses for both of them. Given his preferences, the best that he can do is to defect, whatever Betty does. So he cannot rightly be accused of failing to take his best option by defecting (the same is true of Betty). These points also apply to The Russian Nobleman. Assuming that neither his young nor his old self can settle what he does at both choice points, if he donates to the left in his youth and to the right in his old age, he does what's best in light of his current preferences at each point.[38]

If all of this is right, why are theorists often tempted to think that agents in (at least some of) these cases ought to act in line with the Pattern Perspective? This temptation can be explained, I believe, by the fact that in *repeatable* versions of these cases, which are far more similar to the kinds of situations that we face in ordinary life, universal cooperation is generally in each party's self-interest, even according to the Act Perspective. The fact that a Prisoner's Dilemma is repeated indefinitely allows for the possibility of rewarding those who cooperate with us by cooperating with them in the future, and sanctioning those who defect by refusing to cooperate with them (at least for some time). In general, each of us does better by not being concerned with any particular

[37] In fact, it might be easy for you, at each choice point, to complete a chore, but difficult for you to perform this individually easy task across all of the relevant choice points.

[38] I mentioned in section II that Parfit thinks that it's more objectionable for a moral theory to generate Prisoner's Dilemmas than to give rise to coordination problems. However, in light of the preceding, I actually think that coordination problems are more challenging, since they require us to give up on the initially intuitive thought that for an agent to be accountable for failing to do something, it must have been the case that she ought, all things considered, to have done it.

interaction, but rather by focusing on the indefinitely repeated social interactions characteristic of ordinary life. The best strategy is to be willing to cooperate with anyone initially and to defect against anyone who defects against you (see, e.g. Axelrod 2006).

Provided that adopting this strategy of initial willingness to cooperate with anyone is an option for an agent in a given Prisoner's Dilemma, and that she expects to face an indefinite number of similar situations in the future, the Act Perspective will advise her to adopt it. If everyone takes this option, they'll cooperate with one another, in line with the Pattern Perspective. However, the explanation for why cooperation is justified in these cases is best cashed out in terms of the Act Perspective. You should adopt a strategy of being initially willing to cooperate because you can expect to do better for yourself, based on how others will respond. We don't need to appeal to the fact that, in any given case, universal cooperation is a better pattern for each agent than universal defection.

We can give a similar explanation in intrapersonal, diachronic cases like The Russian Nobleman. If your preferences are going to shift in predictable ways and you realize that these shifts will repeatedly place you in intrapersonal Prisoner's Dilemmas, it's plausible that the best thing for you to do is to adopt a strategy of being willing to cooperate with yourself. For example, if the Russian Nobleman notices that he keeps switching political allegiances, he can adopt a strategy of not donating to either side. If he strays from this strategy and donates (say) to left-wing causes, he can punish his current left-wing self by donating to right-wing causes when his political sympathies shift, and then return, at some point, to his strategy of not donating to anyone.

This example may be somewhat artificial, but plausibly, many of our preference shifts are predictable. Consider various time biases, such as caring much more about some positive experience when it is in the future than when it is in the past, and caring less about a good experience the further it is in the future. For many of us, these preference shifts are entirely predictable, and we do well to recognize when we face these kinds of recurring problems and to find ways to cooperate diachronically with ourselves.

Summing up, taking the Act Perspective doesn't invite mass irresponsibility in Prisoner's Dilemmas. The versions of these cases where it's plausible that we shouldn't take others' behavior as given are cases where they might respond favorably or vindictively towards us in the future, based on what we do now (or where we might respond in this way towards ourselves). We can explain the appropriateness of adopting a generally cooperative strategy in these cases from within the Act Perspective.

6 | SELF-DEFEAT AS A PROBLEM FOR DOMAIN-SPECIFIC THEORIES

I've argued that a normative theory can be self-defeating yet still internally consistent and viable as an advisor and judge. I take these features to be desiderata for *any* normative theory. One possibility that my argument doesn't rule out is that there are certain additional desiderata for particular normative domains such that any theory that's self-defeating is guaranteed to fail to meet some of these domain-specific desiderata and so to be inadequate as a theory of the domain in question.[39]

I grant this possibility. However, self-defeat is most interesting as a domain-general phenomenon. Domain-general, all things considered normative questions have a certain priority over domain-specific questions. Our answers to domain-general questions about what we ought to do

[39] I'd like to thank an anonymous editor for suggesting this possibility.

settle deliberation, whereas our answers to domain-specific questions (e.g. about morality, self-interest, U.S. federal law, chess etc.) only bear on deliberation insofar as they bear on what we ought to do, full stop. With this point in mind, let's consider Parfit's argument that the best *moral* theory cannot be directly collectively self-defeating, and ask whether it tells us anything about the best domain-general normative theory.[40]

According to most views about the nature of morality, Parfit writes, "morality is essentially a collective code—an answer to the question 'How should we *all* act?' An acceptable answer to this question must be acceptable at the collective level" (1984, 106).[41] Granting that Parfit is correct about the essentially collective nature of morality, what ought Andy to do, all things considered, in The Prisoner's Dilemma? Morally speaking, he ought to cooperate. Given his own self-interest, though, he ought to defect. If Andy follows the best moral theory, he'll follow a theory that's not self-defeating. If he follows the best theory of self-interest, he'll follow one that is. Does this difference indicate that Andy should follow morality?

If self-defeat is only a problem for theories of certain normative domains, and not a problem for normative theories as such, then the answer has to be "no." For all that Parfit has told us, the fact that the best theory of self-interest fails as a *moral* theory does not indicate that it fails as a domain-general normative theory. We haven't yet been told anything about what kind of constraints a domain-general normative theory must meet. Additionally, the more that we assume about the moral domain, such that the correct moral theory cannot turn out to be (e.g.) directly collectively self-defeating, the harder it will be to make the case that the best domain-general, all things considered normative theory also has to meet these same constraints.[42]

It's worth noting that Parfit never argues that we should follow morality, rather than self-interest, because only the former is collectively successful. His discussion of collective self-defeat is only meant to cast doubt on certain common-sense assumptions about morality.[43] He offers independent arguments for why moral considerations carry more weight than considerations of self-interest in the best domain-general normative theory.[44]

[40] Recall that for a theory to be directly collectively self-defeating is for there to be situations in which "it is *certain* that, if we all successfully follow T, we will thereby cause the T-given aims *of each* to be worse achieved than they would have been if none of us had successfully followed T" (1984, 55).

[41] Similarly, in *On What Matters: Volume One*, Parfit says that whereas the best theory of individual rationality can be self-defeating, "moral principles or theories are intended to answer questions about what *all* of us ought to do. So such principles clearly fail, or condemn themselves, when they are directly self-defeating at the collective level" (2011, 306).

[42] To bolster this point, it's worth noting that even once we grant that morality is essentially a collective code, a number of further assumptions have to be made to get the result that the best moral theory cannot be directly collectively self-defeating. Preston-Roedder (2014) argues, for instance, that Parfit implicitly assumes that we should only take into account the good and bad consequences of everyone *following* the code. Parfit ignores the possibility that the mere fact that some code *is* the operative code might be good or bad, e.g. by depriving us of moral validation for certain special relationships and projects. Perhaps a more impartial code wouldn't give rise to any Prisoner's Dilemmas, yet it would still partly undermine its own T-given aims simply by being the operative code. Moreover, it's far from obvious how to articulate the collective code thesis in a way that rules out theories that are directly collectively self-defeating, but allows for theories that are merely possibly directly collectively self-defeating in coordination problems (something that Parfit clearly wants to allow (1984, 101)). For helpful discussion of this challenge (though he never puts the problem precisely in these terms), see Southwood 2019.

[43] If morality is essentially a collective code, then we might wonder if Parfit is correct to claim that our common-sense assumptions about (e.g.) the appropriateness of partiality towards one's own children are really assumptions about *morality* at all, and not about some other normative domain. If that were the case, then Parfit's argument against these assumptions would fail on his own terms.

[44] These arguments take up much of parts 2 and 3 of *Reasons and Persons*.

Moreover, one of Parfit's main arguments that's meant to undermine the significance of self-interest relies on the assumption that self-interest requires you to cooperate with yourself in cases of preference shifts such as Russian Nobleman, but not to cooperate with others in interpersonal cases of conflicting preferences such as The Prisoner's Dilemma. Parfit views the fact that theories of self-interest are subject to this diachronic constraint, but not to any collective constraints, as reflecting an irrational bias towards your past and future interests and against the interests of others (1984, 130–32). If Parfit is right, the fact that theories of a certain normative domain are subject to some domain-specific constraints (and not others) may cast doubt on the general significance of the entire domain. This is an additional reason for us to be careful in treating domain-specific constraints as constraints on domain-general normative theories.

7 | CONCLUSION

A self-defeating normative theory has an air of paradox. It seems as though the theory hasn't made up its mind about what it wants from us. It's hardly surprising, then, that so many philosophers reject self-defeating theories. Despite the ubiquity of this move across different domains of normative theorizing, we should stop making it. The fact that a normative theory is self-defeating, as such, gives us no reason to reject it, at least not as a domain-general theory. Each individual response to a situation may be called for, given its alternatives, even though, when we consider the entire set of responses, we see that some other set would have been better. The relevant alternatives on which these verdicts are based differ, so we shouldn't expect our normative theories to always favor the same type of response when answering these different questions.

Once we see that we can judge a response based directly on how it compares to alternative responses (the Act Perspective), or indirectly based on its membership in a set of responses that we compare to other sets (the Pattern Perspective), there is a further question as to which perspective is action-guiding. I argued that only the Act Perspective is action-guiding, because it's the only perspective that takes account of all of an agent's information, including her information about what choices will be freely made at other choice points.

Taking this line invites the charge that we're letting agents off the hook too easily. However, the Pattern Perspective still has a role to play in explaining why agents are blameworthy for some of the choices that they make (or would make). Moreover, once we move away from the somewhat artificial, one-off Prisoner's Dilemmas of game theory to indefinitely repeated interactions, a case can be made that the Act and Pattern Perspectives largely agree about when agents ought to cooperate, and that the Act Perspective offers a better explanation for why each of us ought to be cooperative.

For all that I have said, philosophers who focus on certain normative domains may still be able to find grounds for thinking that theories of those domains ought not to be self-defeating in some way or other. However, the burden is on these philosophers to explain what makes their domains of interest special. They may find that in answering this question, they call into doubt the broader significance of these particular domains.

ACKNOWLEDGEMENTS

Special thanks to Johann Frick, Tom Kelly, and Gideon Rosen for their invaluable feedback on multiple drafts. For additional helpful feedback, I would like to thank Colin Bradley, Liz Harman, Camilo Martinez, Ian McKeachie, Steve White, and an anonymous referee and an anonymous editor for this journal.

ORCID

Samuel Fullhart https://orcid.org/0000-0001-5452-4772

REFERENCES

Andreou, Chrisoula. 2006. Temptation and Deliberation. *Philosophical Studies*, 131, 583–606. https://doi.org/10.1007/s11098-004-8814-x

Axelrod, Robert. 2006. *The Structural Evolution of Morality*. Cambridge University Press.

Bratman, Michael. 1999. Toxin, Temptation, and the Stability of Intention. In Bratman, *Faces of Intention* (pp. 58–90). Cambridge University Press.

Budolfson, Mark. Manuscript. Why Morality and Other Forms of Normativity are Sometimes Dramatically Directly Collectively Self-Defeating.

Carlson, Erik. 1996. Cyclical Preferences and Rational Choice. *Theoria*, 62, 144–160. https://doi.org/10.1111/j.1755-2567.1996.tb00534.x

Christensen, David. 1991. Clever Bookies and Coherent Beliefs. *Philosophical Review*, 100, 229–247. https://doi.org/10.2307/2185301

Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes. 1955. Outlines of a Formal Theory of Value, I. *Philosophy of Science*, 22, 140–160. https://doi.org/10.1086/287412

Dougherty, Michael V. 2011. *Moral Dilemmas in Medieval Thought: From Gratian to Aquinas*. Cambridge University Press.

Dougherty, Tom. 2015. Future-Bias and Practical Reason. *Philosophers' Imprint*, 15(30), 1–16.

Elga, Adam. 2010. Subjective Probabilities Should be Sharp. *Philosopher's Imprint*, 10(5), 1–11.

Estlund, David. 2017. Prime Justice. In M. Weber and K. Vallier (Eds.). *Political Utopias: Contemporary Debates* (pp. 35–56). Oxford University Press.

suppressBEstlund, David.suppressE 2019. *Utopophobia: On the Limits (If Any) Of Political Philosophy*. Princeton University Press.

Fanciullo, James. 2021. The Psychological Basis of Collective Action. *Philosophical Studies*, 178, 427–444. https://doi.org/10.1007/s11098-020-01439-6

Feldman, Fred. 1980. The Principle of Moral Harmony. *The Journal of Philosophy*, 77(3), 166–179. https://doi.org/10.2307/2025668

Gauthier, David. 1994. Assure and Threaten. *Ethics*, 104, 690–721. https://doi.org/10.1093/oso/9780192842992.003.0008

Goldman, Holly S. [Holly M. Smith]. 1978. Doing the Best One Can. In A. I. Goldman and J. Kim (Eds.). *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt* (pp. 185–214). Springer Netherlands.

Hedden, Brian. 2015. Options and Diachronic Tragedy. *Philosophy and Phenomenological Research*, 90, 423–451. https://doi.org/10.1093/acprof:oso/9780198732594.003.0007

Jackson, Frank. 1987. Group Morality. In P. Pettit, R. Sylvan, and J. Norman (Eds.). *Metaphysics and Morality: Essays in Honor of J.J.C.Smart* (pp. 91–110). Blackwell.

suppressBJackson, Frank.suppressE. 2014. Procrastinate Revisited. *Pacific Philosophical Quarterly*, 95, 634–647. https://doi.org/10.1111/papq.12051

Jackson, Frank and Robert Pargetter. 1986. Oughts, Options, and Actualism. *The Philosophical Review*, 95, 233–255. https://doi.org/10.2307/2185591

Kierland, Brian. 2006. Cooperation, "Ought Morally," and Principles of Moral Harmony. *Philosophical Studies*, 128, 381–407. https://doi.org/10.1007/s11098-004-7789-y

Lazar, Seth and Chad Lee-Stronach. 2019. Axiological Absolutism and Risk. *Noûs*, 53, 97–113. https://doi.org/10.1111/nous.12210

Lewis, David. 1999. Why Conditionalize? In *Papers in Metaphysics and Epistemology Vol.* 2 (pp. 403–407). Cambridge University Press.

Marcus, Ruth Barcan. 1980. Moral Dilemmas and Consistency. *The Journal of Philosophy*, 77, 121–136. https://doi.org/10.2307/2025665

Marušić, Berislav. 2015. *Evidence and Agency: Norms of Belief for Promising and Resolving*. Oxford University Press.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press.
Moran, Richard. 2002. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press.
Nefsky, Julia. 2017. How you can Help, without Making a Difference. *Philosophical Studies*, 174, 2743–2767. https://doi.org/10.1007/s11098-016-0808-y
suppressBNefsky, Julia.suppressE. 2019. Collective Harm and the Inefficacy Problem. *Philosophy Compass*, 14(4), 1–17. https://doi.org/10.1111/phc3.12587
Parfit, Derek. 1984 (reprinted with some corrections in 1987). *Reasons and Persons*. Oxford University Press.
suppressBParfit, Derek.suppressE. 1988. Manuscript. What we Together do.
suppressBParfit, Derek.suppressE. 2011. *On What Matters*. Oxford University Press.
Pinkert, Felix. 2015. What if I cannot make a Difference and Know it? *Ethics*, 125, 971–998. https://doi.org/10.1086/680909
Portmore, Douglas W. 2018. Maximalism and Moral Harmony. *Philosophy and Phenomenological Research*, 96, 318–341. https://doi.org/10.1111/phpr.12304
suppressBPortmore, Douglas W.suppressE. 2019. *Opting for the Best: Oughts and Options*. Oxford University Press.
Preston-Roedder, Ryan. 2014. A Better World. *Philosophical Studies*, 168, 629–644. https://doi.org/10.1007/s11098-013-0154-2
Quinn, Warren S. 1990. The Puzzle of the Self-Torturer. *Philosophical Studies*, 59, 79–90. https://doi.org/10.1017/cbo9781139172677.011
Rabinowicz, Wlodek. 1989. Act-Utilitarian Prisoner's Dilemmas. *Theoria*, 55, 1–44. https://doi.org/10.1111/j.1755-2567.1989.tb00720.x
Railton, Peter. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs*, 13, 134–171.
Regan, Donald. 1980. *Utilitarianism and Cooperation*. Oxford University Press.
Soon, Valerie. 2021. An Intrapersonal, Intertemporal Solution to an Interpersonal Dilemma. *Philosophical Studies*. Advance online publication. https://doi.org/10.1007/s11098-021-01604-5
Southwood, Nicholas. 2019. Contractualism for us as we are. *Philosophy and Phenomenological Research*, 99, 3, 529–547. https://doi.org/10.1111/phpr.12500
Sullivan, Meghan. 2018. *Time Biases: A Theory of Rational Planning and Personal Persistence*. Oxford University Press.
Timmerman, Travis and Yishai Cohen. 2016. Moral Obligations: Actualist, Possibilist, or Hybridist? *Australasian Journal of Philosophy*, 94, 672–686. https://doi.org/10.1080/00048402.2016.1140789
suppressBTimmerman, Travis and Yishai CohensuppressE. 2020. Actualism and Possibilism in Ethics. In E. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/actual-ism-possibilism-ethics/
Tuck, Richard. 2008. *Free Riding*. Harvard University Press.
White, Stephen J. 2015. The Problem of Self-Torture: What's Being Done? *Philosophy and Phenomenological Research*, 94, 584–605. https://doi.org/10.1111/phpr.12261
Wiland, Eric. 2007. How Indirect can Indirect Utilitarianism be? *Philosophy and Phenomenological Research*, 74, 275–301. https://doi.org/10.1111/j.1933-1592.2007.00018.x
Woodard, Christopher. 2008. *Reasons, Patterns, and Cooperation*. Routledge.
Wu, Patrick. 2022. Aggregation and Reductio. *Ethics*, 132, 508–525. https://doi.org/10.1086/716868
Zimmerman, Michael. 1996. *The Concept of Moral Obligation*. Cambridge University Press.

How to cite this article: Fullhart, S. (2023). Embracing self-defeat in normative theory. *Philosophy and Phenomenological Research*, 1–22. https://doi.org/10.1111/phpr.13033