An intrapersonal, intertemporal solution to an interpersonal dilemma

[Penultimate version, 1/9/2021. Please cite the final version, forthcoming in *Philosophical Studies*.]

Valerie Soon
Duke University
vys6@duke.edu

Abstract

It is commonly accepted that what we ought to do collectively does not imply anything about what each of us ought to do individually. According to this line of reasoning, if cooperating will make no difference to an outcome, then you are not morally required to do it. And if cooperating will be personally costly to you as well, this is an even stronger reason to not do it. However, this reasoning results in a self-defeating, yet entirely predictable outcome. If everyone is rational, they will not cooperate, resulting in an aggregate outcome that is devastating for everyone. This dismal analysis explains why climate change and other collective action problems are so difficult to ameliorate. The goal of this paper is to provide a different, exploratory framework for thinking about individual reasons for action in collective action problems. I argue that the concept of commitment gives us a new perspective on collective action problems. Once we take the structure of commitment into account, this activates requirements of diachronic rationality that give individuals instrumental reasons to cooperate in collective action problems.

1. The structure of collective action problems

Climate change is caused by anthropogenic carbon emissions that each of us generates in the course of our daily lives. It is too late to completely reverse climate change, so the best thing we can now do is to slow or stop its progress. To do so, we need to collectively cut back on our carbon emissions. While it is clear that climate change could be mitigated by a sufficient number of individuals cutting back on their carbon emissions, it is equally clear that no individual action makes a difference to the progression of climate change.

Climate change is just one pressing example of a collective action problem that has the structure of the "tragedy of the commons" (Hardin 1968). While there are several types of

collective action problems,[1] the ones of interest here are situations in which many individual actions create a harmful outcome in the aggregate, but each individual action seems to make no difference to the outcome. Therefore, there seems to be no reason for the individual to "cooperate": to take the action that would be part of the set of actions that bring about the desired outcome. Voting in national elections, participation in exploitative supply chains, and complicity with harmful social norms also have this structure. It may be true that an individual action can "make a difference" by changing the description of the outcome on a fine-grained level, but an individual action doesn't change the outcome. Your vote could reduce the loss margin of your favored candidate by one vote, but your candidate still loses. Therefore, many individuals rationally defect rather than cooperate, even if they want their party to win, are against exploitation, or reject harmful social norms. We defect because cooperating is not only apparently inefficacious, but also personally costly, adding a further reason against cooperating. Collective harm is a predictable result of this reasoning and could be avoided if enough people decided to cooperate. But what reason does any individual have to cooperate, when an individual act of cooperation makes no difference to the outcome? Julia Nefsky (2019) calls this the *inefficacy problem.*

As Nefsky and other moral philosophers have argued, the inefficacy problem makes it extremely difficult to find a moral reason for action in collective action problems. This debate parallels a longstanding debate in rational choice theory, where the consensus view is that no instrumental reasons to cooperate can be derived from the preference to make a difference in collective action problems. If this is the only preference under consideration, the dominant strategy is to defect rather than to cooperate. Therefore, rational choice theorists have sought to

[1] E.g. Coordination problems such as the Assurance Game or Chicken.

explain cooperation by appealing to preferences that are only indirectly related to the outcome, such as expressive preferences.

My strategy is a bit different from both these camps. The goal of this paper is to excavate an instrumental, outcome-based reason for action in collective impact cases by appealing to diachronic rationality – rationality over time, especially with respect to our commitments and values. The core idea behind my argument is this: in daily life, each of us has commitments that require many individually ineffective actions to be realized. It is often rational to "defect" at a time, or to put off taking one individual action. If we do this often enough, we will fail to realize the relevant commitment. However, commitments bind us to "cooperate" even when this seems irrational at a time, because we know that defecting too often will undermine our ability to follow through on our commitments. In this paper, I argue that this strategy, quotidian as it is in daily life, can shed light on our reasons to cooperate in collective situations. Throughout this paper, I will remain agnostic as to whether moral reasons can be provided for action in collective impact situations.

The rest of this paper proceeds as follows. In Section 2, I critically sketch out different approaches to the inefficacy problem and the collective action problem that it derives from; I argue that calls toward purely structural change or political advocacy merely give rise to another iteration of the inefficacy problem. In Section 3, I draw on Michael Bratman's planning framework to argue that switching to a diachronic perspective can give us outcome-based, instrumental reasons to cooperate in collective action problems, in spite of the pull of the inefficacy problem. This allows us to bypass the expected utility debate. Section 4 argues that these reasons are relative to each individual's commitment to the good. Section 5 concludes.

Before launching into the main argument, it will be good to make clear my assumptions, methodological commitments, and the scope of my target. I will assume that diachronic rationality is a distinct concept from synchronic rationality, which concerns rationality at a time, and that there is such a thing as a subjective reason. I assume, along with much of the collective action literature, that collectives are constituted by individuals, so collective action requires individual action;[2] hence the focus on individual reasons for action. I also assume, more controversially, that intentions, particularly commitments, carry with them certain norms of coherence, consistency, and stability.

Finally, my arguments are targeted toward a class of individuals with a particular psychology, not to everyone. The Ordinary Person is someone who is rationally compelled by the inefficacy problem towards inaction. Unlike someone who does not care at all about collective harm, the Ordinary Person is committed to the good. But she is not someone with an activist-like mindset who acts on deontological commitments, nor does she have an overwhelming desire to disregard instrumental calculations in favor of virtuous action. The inefficacy problem is truly a *problem* for this person; she would like to act but finds no outcome-based reason to do so. Thus, she is the person whom we need to provide reasons for. Even though my arguments are limited in scope, I think that the Ordinary Person is a challenging enough target. She is familiar to many of us.

2. The inefficacy problem

[2] This ontological position does not entail methodological individualism, which holds that all human activity is best *explained* in terms of the activity of individuals (Elster 1989). Methodological holists usually also accept the claim that all human activity is *ontologically* constituted by the activity of human individuals.

The inefficacy problem lies at the heart of collective action problems. According to the inefficacy problem, "if acting in the relevant way won't make a difference, it's unclear why it would be wrong. Each individual can argue, 'things will be just as bad whether or not I act in this way, so there's no point in doing otherwise'" (Nefsky 2019, 2). Acting in the relevant way ("cooperating", for short) doesn't make a difference because it is neither necessary nor sufficient for bringing about the desired outcome.

The inefficacy problem is starkest in cases where an individual action does not cause any small-scale, localized harm by itself, so we cannot bypass the inefficacy problem by appealing to these small-scale harms to deliver moral reasons against action. For example, consider segregation: if a white, wealthy family moves from the city to the suburbs, they do not harm anyone. But when enough of these families do so, this is known as "white flight", which drains cities of economic resources and perpetuates racial segregation (Massey and Denton 1993, Kruse 2013).

Let's distinguish between two versions of the inefficacy problem. On the weak version of the inefficacy problem, individuals are not morally required to cooperate on a first-order level if doing so won't make a difference. No one is required to stop driving if doing so won't make a difference. Assuming that there is such a thing as supererogatory action, it could be morally good or virtuous of individuals to cooperate, but the key point is that we are not *required* to (Sinnott-Armstrong 2005, with Kingston 2018). Rather, what we are required to do is to cooperate on a second-order level, by engaging in political advocacy to push our government to act against collective harms (Sinnott-Armstrong 2005, 2018 p.185). For example, we ought to push our governments to regulate gas-guzzling SUVs, but we need not refrain from joyriding in these SUVs. Call this solution *Political-Not-Personal (PNP).*

However, PNP is vulnerable to the inefficacy problem as well.[3] It is not clear why we are morally required to take political action if individual political action doesn't make a difference and is also personally costly. It is not realistic to expect any individual political action to make a difference to an outcome such as climate change; for this to happen takes a great deal of luck and is exceedingly rare, if it ever happens.[4] Given the difficulty of this problem, various strategies to make sense of voting, protesting, and organizing have appealed to fact that political action aims to secure some form of private benefit to the individual, rather than to its direct efficacy in securing a public good. For example, political action may help activists gain personal status (Tullock 1971), help those who are unhappy with a social norm satisfy a direct preference for conditional cooperation (Bicchieri 2005), or function as a means for expressing one's political preferences (Lomasky & Brennan 1993). Or political action may simply be a way to act in accordance with one's moral beliefs, whether these be grounded in virtue or deontological principles. None of these strategies appeal to efficacy itself.

Moreover, the inefficacy problem is heightened by the costs of political action. Take voting as a simple instance of political action. The costs of voting can range from the relatively trivial, such as standing in line to cast a vote; to the more expensive, such as deeply researching candidates' policy positions and the relevant social science so as to be able to cast an informed vote; to extremely serious ones such as dangers to one's physical security in situations where voter intimidation is a problem. In addition to these direct costs of political action, there are also significant opportunity costs, such as the time and effort one could spend on one's leisure

[3] Nefsky (2019) raises this point in footnote 8.
[4] By "luck", I mean that the individual must take the right sort of action, have the right sort of personality, and ultimately, being in the right place at the right time, among other factors. For example, Greta Thunberg is one person who has had an outsize impact as a climate change activist, but even with her prominence, carbon emissions continue unabated.

pursuits, personal relationships, or career. If the inefficacy problem renders even an action as simple as voting irrational, then it is not clear what efficacy-based reason could be provided for more aggressive forms of political advocacy, such as direct action against environmental injustice. Therefore, for those who are focused on the efficacy of action, it is not clear what the justification is for a moral or political obligation to undertake political action. PNP cannot be justified by the efficacy of political action.

PNP also fails on a strong version of the inefficacy problem. On the strong version, you not only have no moral requirement to act; you also have no moral reason to, supererogatory reasons included. As Nefsky writes, "When one says 'but it won't make any difference,' more than just saying 'it doesn't seem that I am obligated to act in that way', one is saying 'there doesn't seem to be any point in acting that way.' Doing so, in this light, looks like a mere waste" (Nefsky 2017, 2744-2745).

The inefficacy problem applies in both threshold and non-threshold cases. Let's first look at threshold cases. In threshold cases, once a threshold is reached and exceeded, a significant harm results. To avoid crossing the threshold, a certain number of people are required to cooperate. For example, food supply chains are thought to fall under the umbrella of threshold cases (Kagan 2011, Budolfson 2018). We might abhor the suffering caused to animals in factory farms, but there is no clear efficacy-based reason to cooperate (i.e. refrain from buying factory-farmed meat). If not enough people cooperate, then your cooperation will be futile. If enough other people cooperate, then your cooperation is superfluous, so there is also no reason to cooperate (Nefsky 2019). Therefore, the dominant strategy is non-cooperation.

In many collective action problems, we will not be operating under anything like the degree of certainty implied in the claim that acting "*won't* [italics mine] make any difference".

Most of the time, we are confronted with situations in which we are simply not sure whether doing so will or won't make a difference. As such, Shelley Kagan (2011) and other consequentialists, such as Alastair Norcross (2004), Peter Singer (1980), and John Broome (2019), have appealed to expected utility in threshold cases. As a normative theory of rational choice, expected utility theory tells us to choose the act with the highest expected utility. Kagan argues that in threshold cases, a small probability of being the triggering act, multiplied by the consequences of the outcome, will always deliver negative expected utility, so consequentialism can straightforwardly condemn the act on the grounds of its negative expected utility (Kagan 2011, 119-120). Even if one's action is unlikely to make a difference, its negative expected utility tells us not to do it.

Nefsky (2011) and Mark Budolfson (2018) have come out with arguments against expected utility that are, to my mind, convincing. Both argue that we cannot a priori conclude that expected utility will always come out negative in every threshold case. Nefsky says that "Whether it does or not depends on the probabilities and on the goodness or badness of the relevant consequences" (2011, 369). Budolfson makes an even stronger claim. He argues that due to empirical facts about supply chains in industrial agriculture, it will *rarely* be the case that expected utility shakes out in favor of action in that sort of case. The mistake that Singer, Norcross, and Kagan make is that they assume that expected utility of an action is equal to the average effects of all similar actions. However, it is more likely that the expected utility of cooperating approaches zero, so expected utility provides weak consequentialist grounds for cooperating.

Once we also consider the instrumental costs of cooperation, Nefsky and Budolfson's skepticism about the ability of expected utility to ground reasons to cooperate is strengthened.

Assuming an individual has a preference to help bring about the desired outcome, she can expect that her action will most likely not do so, and that its utility will be outweighed by the costs of action. In sum, expected utility cannot provide a general solution to the inefficacy problem, especially when we expand the scope of the relevant consequences to personal costs.

Expected utility has even more difficulty in handling non-threshold cases, which are cases where there is no tipping point at which an outcome decisively occurs. Non-threshold cases resemble Sorites paradoxes, in which predicates with vague boundaries generate difference-making problems. Consider the harmless torturer (Parfit 1984) and the puzzle of the self-torturer (Quinn 1990). In the harmless torturer case, a thousand torturers turn a switch on some instrument once. Each turn increases a victim's pain imperceptibly, but the aggregate effect of a thousand turns is that each victim ends up in severe pain. However, "none of the torturers makes any victim's pain perceptibly worse" (Parfit 1984, 80). At no point does expected utility tell each of the torturers that they should not turn the dial, so none of them act wrongly.

Quinn's puzzle of the self-torturer is similar, but its point is to show that an agent can make a series of rational decisions that land her in an unwanted outcome due to the intransitive structure of her preferences. The reasons involved in this situation are entirely instrumental. The self-torturer is hooked up to a medical device that increases her pain level by one imperceptible increment at each turn of the dial. At each turn of the dial, she receives $10,000. Because any individual increase in pain is imperceptible, it is rational for her to turn the dial to receive the $10,000. But the self-torturer eventually ends up in a state of severe pain, which she disprefers to the total amount of money she's gained, due to her intransitive preferences. Orthodox rational choice theory (RCT) tells us that the self-torturer is irrational because her preferences are intransitive, so it is no surprise that expected utility leads her to an unwanted outcome. However,

the self-torturer can also be seen as presenting a difficulty for orthodox RCT, since it does not seem irrational for her to trade an imperceptible increase in pain for a perceptible increase in money (Andreou 2006). These non-threshold cases are more difficult than threshold cases, for expected utility issues a verdict in favor of the action (i.e. non-cooperation) that, when iterated enough times, will turn out to lead to a harmful aggregate outcome.

To summarize so far, expected utility is not a promising route for providing individual reasons for cooperation, as Nefsky and Budolfson have argued. But a moral or political obligation to advocate for governmental action (in the cases where this would be an effective intervention) is vulnerable to the inefficacy problem as well, so Political-Not-Personal, as articulated by Sinnott-Armstrong and Kingston, is at least inconsistent. Moreover, both moral and instrumental reasons appear to counsel us against cooperating. Thus, the inefficacy problem is a serious problem on multiple dimensions.

Much of the debate on the inefficacy problem has focused on moral rather than instrumental reasons. The theoretical reason for this is that the inefficacy problem poses a problem for consequentialism. In addition, there are also practical explanations as to why we should lean on moral rather than instrumental reasons. As the discussions of the torturer cases show, it is instrumental reasoning that seems responsible for collective harm. Each person can argue that "it's not my fault" (Sinnott-Armstrong 2005) because their individual contribution makes no difference to the outcome, and that it would be personally costly for them to change their behavior. In fact, we often hear Ordinary People make a version of this argument. Since instrumental reasons create the problem, we might think that instrumental reasons cannot get us out of it. That is the negative explanation – why *not* instrumental reasons. Therefore, focusing on instrumental reasons sharpens the problem.

The positive explanation – *why* moral reasons are at the focus of this debate– requires making some conjectures about the metaethical assumptions framing the debate. One possible explanation is that moral reasons are assumed to have universal normative force; they are supposed to swamp instrumental reasons. If agent-relative, instrumental reasons create the problem, we need agent-neutral reasons, such as those provided by morality, to get us out of it. But if the above interlocutors are right about the inefficacy problem, as I think they are, morality doesn't seem to be fertile ground to harvest reasons for cooperating in collective action problems. Because the inefficacy problem is so difficult, Nefsky (2017) has turned her attention to undermining the importance of difference-making for grounding reasons for action.

The aim of this paper is to explore territory that has been left within the space of reasons. So in what follows, I will take a different strategy, focusing instead on temporally extended instrumental reasons. I will argue that the Ordinary Person has instrumental reasons, given by diachronic rationality and self-governance, for cooperating in collective action problems. This strategy allows us to bypass the debates about expected utility to reach a conclusion that applies even to the difficult non-threshold cases.

3. An intrapersonal, intertemporal solution

The argument that follows builds on the core intuition that collective action failures and intrapersonal, intertemporal failures share the same overall structure: no one act makes a difference, but many acts together do, and enough instrumentally rational failures to cooperate will result in the failure to achieve the overall goal. But we have a good solution to this in intrapersonal cases, such as long-term projects. In such cases, we adopt commitments as a device to get over the inefficacy problem. I will argue that commitment can also help us get over the

inefficacy problem in the collective action case. Making sense of commitment requires taking a diachronic perspective on rationality.

First, let's introduce two forms of instrumental rationality, synchronic and diachronic. An agent is instrumentally rational if and only if she intends the necessary means to her intended ends. Synchronic rationality concerns rationality at a time, whereas diachronic rationality concerns rationality over time (Bratman 2007), or constraints on intertemporal combinations of intentional states. Both forms of rationality require at least *means-end coherence* and *consistency*. To be coherent, the agent should intend what they believe to be the necessary means to their intended end. To be consistent, the agent should not intend A and B if they believe that A and B are not co-possible. In addition to the norms of coherence and consistency, diachronic rationality also includes the norm of *stability of intention*: your commitments shouldn't change for arbitrary reasons (Bratman 2012). The connection between belief and action will become important later.

Synchronic and diachronic rationality can conflict, though not necessarily. Sometimes, we make choices that are rational at each point in time, but a series of such synchronically rational decisions leads to a diachronically irrational outcome. The conflict arises because of what each temporal perspective on rationality tells us to do. Synchronic rationality tells us that we should maximize expected utility at a given time. This is what the self-torturer does by opting to turn up the dial. Diachronic rationality rejects this claim; rather, we should do whatever is necessary for us to realize our long-term commitments, even if this sometimes requires not maximizing expected utility at a time. For an example of diachronic rationality in operation, consider Ulysses contracts. These precommitment devices help us "bind ourselves to the mast" by overriding present desires, thereby preventing us from succumbing to temptations that would

lead us astray from our commitments. They can be as quotidian as self-imposed 25-minute "Pomodoro" time blocks, during which you are not supposed to do anything but your prescribed task no matter what, or they can be as weighty as advance directives that allow people to spell out end-of-life decisions in case of dementia.

The conflict between synchronic and diachronic rationality is apparent in both threshold and non-threshold cases. Consider the following non-threshold case, *Fitness*. Suppose Hera sets herself the goal of being fit by the end of the year. This is a vague goal: there is clearly a difference between someone who is fit and someone who is not, but there is no precise threshold that we can point to to identify this difference.[5] To be fit, she must work out with some intensity most days of the week. But also suppose that occasionally missing a workout or two won't undermine her goal; it won't make a difference to her fitness level. And perhaps something more pressing comes up, or she simply doesn't have the desire to work out that day. If, at every choice point, Hera thinks, "I can miss this workout, because doing so won't make a difference to my fitness goal," she will not achieve her goal. There is something irrational about this chain of reasoning—after all, she is not taking any steps towards my long-term goal. There is also something extremely tempting and reasonable about it;[6] after all, no synchronically rational choice undermines her goal. Hence the conflict between the synchronic and diachronic.

Now consider a threshold case, *Running*. Hera sets a goal of running a 6-minute mile in one year's time, to be measured during a race. She either achieves it or she does not. She currently runs a 10-minute mile. In order to achieve her goal, she needs to run at least 70 times

[5] There is some debate about the possibility of a vague goal. I assume that we can aim for vague goals: to be fitter, morally better, etc. See Tenenbaum and Raffman (2012) for an argument that they are possible–we aim for them all the time–and that a theory of rational choice needs to make sense of this.

[6] The apparent reasonableness of this chain of reasoning underlies the temptation of procrastination (Andreou and White 2010).

during the year.[7] But Hera is a busy and somewhat distractable person, and oftentimes, she would rather skip a run to do something else. Can synchronic rationality tell Hera to run enough to achieve her goal?[8] It may or it may not. Earlier in the year, it will often be synchronically rational for her to skip runs whenever she is busy, assuming that the value of the alternative activity is sufficiently high at the time, as long as she is still able to complete at least 70 runs. Perhaps she is involved with travel, or taking care of her grandparents, or giving a series of talks. It doesn't really matter for achieving her goal that she skips her run on these busy days. It will maximize her expected utility on each day to do so, as long as the expected costs of skipping a run on a given day earlier in the year are lower than the expected costs of skipping a run on at least one day late in the year.

Predictably, however, synchronic rationality can sometimes lead Hera into a last-minute push during the last 70 days of the year to make up for lost time. Suppose Hera has to finish an important project in 9 months; but she need only reach her 6-minute mile goal in 12 months. The opportunity costs of running before the project is due are higher than after, so synchronic rationality can lead Hera to that last-minute scramble.[9] At that point, synchronic rationality will tell her that she must run each day in order to achieve her goal, even if opportunity costs are high, as long as the value of achieving her goal is higher than the opportunity costs of running. This is a bad situation to be in. During these last 70 days, there may be a day in which the opportunity costs of working out are so high (e.g. there is a personal emergency that must be attended to, she falls sick) that she is forced to skip a run, thereby eliminating the possibility of

[7] Obviously, this case greatly simplifies how exercise works.
[8] I thank an anonymous reviewer for pressing this objection and suggesting this example.
[9] Thanks to an anonymous reviewer for suggesting this case.

achieving her goal.[10] Therefore, synchronic rationality can lead Hera to fail to achieve her goal through a series of individually expected-utility-maximizing decisions.

If Hera is diachronically rational, she will avoid putting herself in such a situation. She will look forward in time and plan accordingly, incorporating into any present decision the expectation that opportunity costs will likely be higher in the future if she skips today's run. Even if it maximizes her expected utility at one point in time to skip a run, she will not always allow that calculation to justify skipping. Her commitment structures her decisions. Thus, she will run more consistently throughout the entire year rather than leaving everything to the last 70 days.

In both the threshold and non-threshold cases, the conflict between synchronic and diachronic rationality means that expected utility will not always tell us to "cooperate" in service of the overarching goal. It may plump in favor of cooperation, but it just as easily may not, depending on the contingent structure of preferences at each decision point. And what we are seeking is a theory that tells us to cooperate at a necessary number of decision points. Diachronic rationality, which tells us to structure our actions based on our commitments, doesn't allow us to fall into the trap of potentially self-defeating synchronic utility-maximization.

These individual cases of temporal conflict share the same structure as collective action problems. In both types of cases, one individual action makes no difference, but noncooperation for that reason will predictably lead to a suboptimal outcome. But commitment tells us to

[10] I am simplifying the cases such that the law of diminishing returns does not crop up, since that is not a feature of the inefficacy problem. Nevertheless, it's worth addressing what happens when we do take diminishing returns into account. One might object that most activity tends to follow the law of diminishing returns, so synchronic rationality will tell Hera to run early on when the marginal benefits are high. Assume that the activity in question does follow this law. Here, synchronic rationality is still insufficient—there is no guarantee that it will tell Hera to run when marginal benefits begin to level off. At that point in the curve, it's possible that the marginal costs of running will exceed the marginal benefits at each point, even if continuing to run is necessary to achieve her overarching goal. I thank an anonymous reviewer for prompting me to address this point.

cooperate in the service of our long-term interests in the individual case. As I will show, commitment can perform the same role in collective cases. In fact, it already does in some fairly quotidian, artificial collective action problems, in which many of us ignore the pull of the inefficacy problem. Specifically, we don't buy this argument in the case of organized collective activities such as team sports; we think that we have reasons for action in such situations, despite the fact that the inefficacy problem applies. I will use this observation as an anchoring point to build an argument from coherence in favor of an instrumental reason for action.

Consider an intrapersonal, intertemporal dilemma built into an interpersonal dilemma**,** *Rowing* and *Gymnastics*. *Rowing* is a threshold case:

> You are in a boat race with a 9-person crew. The race has a highly unequal prize structure: you either win and receive $10,000, or you don't and receive nothing. You strongly desire to win. You are slightly behind the next crew, and from your limited vantage point, they seem to be pulling ahead with 500 meters to go. There is still a small chance that you might win. You are exhausted and in pain, and your strongest desire at this moment is to pull just a tiny bit less hard to alleviate your pain. After all, there are seven other rowers in your boat. Dialing down your effort would barely be perceptible, if at all. You have no assurance that everyone else is pulling as hard as possible. You know that maximum effort over the length of the race by at least four other rowers will be required for you to win, and if at least six other rowers are putting out maximum effort, then you will definitely win.

The inefficacy problem applies here. Even if it is synchronically rational for the boat as a whole to put in maximal effort at each stroke, it doesn't follow that it is synchronically rational for each individual to do so. As with the *Running* and *Fitness* cases, the cost of putting in

maximum effort at a time is high, and the probability that that particular stroke will lead to the win is exceedingly low. The inefficacy of each stroke is magnified by the necessity of others also working hard: there are eight rowers in the boat, and each typically takes hundreds of strokes over the course of a race. If fewer than four or six rowers are cooperating, then your effort will be wasted relative to either of the two possible good outcomes. If more than four or six rowers are cooperating, then your effort is not required. So, in either of these scenarios, it seems rational for you to not put in maximal effort. That is, the chance that the expected utility of a stroke will turn out negative is higher than not in either scenario, so one should not exert maximal effort. Synchronic rationality therefore can tell each rower that it is fine to keep the pressure off on each stroke. Noncooperation seems like the dominant strategy, giving rise to an interpersonal dilemma. Yet following this prescription will put everyone in a last-ditch situation similar to the one Hera faces in *Running* and will significantly raise the likelihood of a loss.

Next, consider a non-threshold case, *Gymnastics*:

> You are on a team of eight gymnasts competing at a meet. For your team to win, each gymnast needs to acquire a certain number of points, and to do so, each of you will have to execute your routine successfully. But what counts as success is vague; no particular action results in success. Success is evaluated holistically and requires that the judges believe that the routine as a whole has been done well. To do well, you're required to execute a series of strenuous, potentially dangerous moves. The danger of a move increases with the effort, or power, put into it. You have no assurance that anyone else on your team will put in maximum effort to perform well. Maximum effort by at least four other gymnasts will be required for

you to win, and if at least six other gymnasts are putting out maximum effort, then you will definitely win.[11]

Compared to *Rowing*, the inefficacy problem is heightened in *Gymnastics*. While in *Rowing*, it is merely possible or likely that expected utility will turn out negative, the vague structure of success in *Gymnastics* ensures that no move will make a difference to the outcome. But this narrowly synchronic form of reasoning seems inappropriate in both *Rowing* and *Gymnastics*. In collective activities such as sports and other group activities, most people don't act in the way licensed by the inefficacy problem.[12] Most people will put in maximum effort at every choice point. We do so because we take commitments seriously. Participation in a collective endeavor entails committing to a common goal, and commitment overrides the synchronic, instrumentalist reasoning that would otherwise give rise to a collective action problem.

In these collective cases, even if an individual is certain or nearly certain that her individual cooperative action (or even all of her individual cooperative actions) will not make a difference to the outcome, commitment still requires her to cooperate. To see why, consider the epistemic situation of the agent. The agent has committed to a collective goal: a goal that is achievable only with collective action. And she must believe that it is achievable only with collective action, in order for the inefficacy problem to get a hold on her. If she did not believe this, then she would not be able to say that her individual cooperation doesn't matter, because it doesn't matter only if others' cooperation is necessary or sufficient to achieve the goal. So, if the agent believes that the goal can only be achieved through collective action, she knows that the expected utility of her action will either be negative or zero (in the non-threshold case) or be very

[11] This case is constructed with a layperson's understanding of gymnastics, with apologies to those who know better.
[12] See Nguyen (2019) for an overview of the implications of game-playing for practical rationality.

low to negative (in the threshold case). Either way, expected utility tells her that she shouldn't cooperate, because doing so will certainly or nearly certainly make no difference to the outcome. Should she follow expected utility's dictates?

She should not. If she is committed to her goal, she must take the actions necessary to achieve it. Here, the relevant necessary action is the *set* of individual actions that together bring about that outcome. The agent's action may not *itself* be necessary within that set, but because commitment implies action, she is required to participate in that set of actions, even if her action itself ends up being inefficacious. And because she has committed, the question "why should I, in particular, cooperate, when doing so doesn't make a difference?" cannot arise for this agent. This question is puzzling, for it is like saying: you are committed to your team winning, you know that this depends on the cooperation of everyone else, your action by itself is certain or not to make a difference, or is very unlikely to make a difference, and your action alone doesn't matter–therefore you'll just skulk around on the sidelines. Clearly, this line of reasoning is incoherent. An agent who employs this line of reasoning does not seem genuinely committed, and if she is committed, she has not thought through what commitment requires.

There are a few disanalogies to tidy up between the artificial cases and the real collective action problems. First, the cases illustrate outcomes that require a series of actions to achieve, but real cases are sometimes one-shot. It is a short step to generalize the argument to one-shot situations. In one-shot situations, it is imperative that one cooperate, or one has failed to participate in the set of actions necessary to achieve the goal.

Second, there is an importance difference between the real-world cases and artificial sports cases. The real cases are not organized, artificial activities with arbitrary rules. We find ourselves in collective action problems not of our own volition, and not because we are

deliberately acting in concert with others, but due to social, political, and economic structures. This disanalogy naturally raises the question: even if commitment gives us a reason for action in artificial collective action problems, why does this reason apply in general collective action problems?

Return to the Ordinary Person, whose psychology I described in Section 1. The Ordinary person is the target of these arguments. She is sincerely committed towards bringing about the good. It's just that the inefficacy problem blocks her from taking action. When the agent makes a commitment to bringing about the good, her situation is similar to that of athletes who have committed to winning as part of a team. Thus, she enters into a *Rowing*- or *Gymnastics*-like situation with actors who have made a similar commitment, even if action is diffuse across time and space, and even if she never forms actual relationships with the other actors. In these situations, we ought to act in spite of the inefficacy problem.

In what cases does commitment require cooperation? My account is similar to Nefsky's (2017, 2760-2764), who argues that an individually inefficacious action is required as long as the outcome that it can help to bring about is possible. There is a distinction between two types of situations where individual cooperation is certainly or almost-certainly inefficacious: cases in which the desired outcome is certain or impossible, and cases in which the outcome has yet to be determined and is possible. Commitment does not require (knowingly) futile cooperation, only cooperation in cases where the desired outcome is a live possibility. In cases where the outcome is certain, and it is therefore certain that each individual act of cooperation will not make a difference to the outcome, the individual is not required to cooperate. In *Rowing*, for example, if the other boat is just about to cross the finish line, but your boat is still 200 meters away from it, then the outcome is settled. Nothing short of an act of God would help you win. In this case, you

are no longer required to exert maximum effort. Therefore, commitment requires cooperation only if it is uncertain whether the outcome can be attained, and even if it is certain or nearly certain that each individual act of cooperation will not make a difference to the outcome.

My account can be thought of as a synthesis and development of recent work in practical rationality that challenges, or at least sidesteps, the prescriptions of orthodox RCT. Sergio Tenenbaum and Diana Raffman (2012) argue that to make sense of vague projects, an extended temporal perspective is needed to overcome the prescriptions of orthodox rational choice theory, which restricts our pursuits to well-defined ends. However, they argue that planning is not necessary for us to carry out vague projects; we simply need to carry out a sufficient number of the relevant actions. This account does not consider the opportunity costs associated with a lack of planning. I show that the concept of commitment requires us to take an extended perspective to both vague and well-defined ends, and its implicit planning helps us act prudently to ensure that the desired outcome will be achieved. Margaret Gilbert (2006) argues that "joint commitment", wherein two or more parties collectively intend to perform an action as a body, can provide sufficient reason to cooperate in collective action problems. On my account, an individual commitment is enough to provide this reason, and it can do so even if individuals do not intend to act together. Finally, Nefsky (2017) argues that one has a moral reason to cooperate if doing so helps to bring about a good outcome, even if one's action doesn't make a difference, as long as one's action makes a non-superfluous causal contribution. My account buttresses Nefsky's view by showing that moral reasons for cooperation in collective action problems are structurally similar to, and continuous with, our purely instrumental reasons for action in non-moral situations. Commitment allows us to get the same result without any recourse to moral

language. It also gives us the resources to overcome the instrumental, self-regarding calculations that stand in the way of cooperating, even when we know we have moral reasons to do so.

4. Why stick to your commitments?

I have argued that commitment implies action, regardless of the efficacy of each individual action towards the goal. We can use the resources of self-governance to strengthen this instrumental reason for action. A self-governing agent is governed by her commitments and values, not by her occurrent desires (Bratman 2003). Harry Frankfurt's (1971) unwilling addict is the classic example of a non-self-governing agent – one who acts on his first-order desire to smoke, but against his second-order desire (perhaps a commitment) to cease smoking. If self-governance is intrinsically valuable, as Bratman (2018) has argued, it gives each person a reason to act in accordance with their commitment, regardless of how strong a commitment is. However, skeptics such as Niko Kolodny (2008) have pushed back against the intrinsic value of self-governance. I do not have the space to defend the intrinsic value of self-governance in this paper, nor do I want my account to hang on the resolution to this problem. Nor do I want to further restrict the scope of this reason to act, so that it only applies to those who have an intrinsic desire for self-governance. To leverage self-governance in favor of action, I will therefore appeal to the personal value to the Ordinary Person of being a person who is committed to the bringing about the good.

In the above artificial cases, we appealed to the structure of commitment to derive a reason for cooperation in spite of the inefficacy problem. But the reason for cooperation seems overriding only if the agent is isolated from other conflicting commitments, as happens in games. In real life, however, the Ordinary Person frequently finds conflicts between the commitment to

the good and other commitments. Return briefly to the examples of climate change and segregation that we began with. Not all greenhouse gas emissions are caused by frivolous pursuits such as joyguzzling (Sinnott-Armstrong 2005); perhaps one needs to frequently take long-distance flights to visit loved ones across the country because one is committed to maintaining these relationships. Not all segregation is caused by racial animus or indifference to racial justice; much of it stems from the desire to give a good life to one's children by moving to a better neighborhood with more resources. Call these "individual commitments", for lack of a better term. When commitments conflict, we need to give up one thing in favor of another, even if we do so regrettably. This raises the question: why prioritize the commitment to this particular good – the commitment to alleviating collective harm – rather than the commitment to other important projects and relationships?

To justify the priority of this commitment, we need to delve into the requirements of self-governance. Self-governance requires that an agent adhere to the practical norms of consistency, coherence, and stability of intention with respect to her commitments. I will address each of these in turn. Begin with the norm of consistency, which gives us a weak response to this problem. According to *consistency*, a rational agent should not intend A and B if she believes that A and B are not co-possible (Bratman 2012, 73). So she should simply drop either A or B. However, a rational agent with sufficient information about the collective action problem should, presumably, also have knowledge of the inefficacy problem. The inefficacy problem opens room for the possibility that one can intend to commit to the good, yet also make other seemingly conflicting commitments. After all, if my act of cooperation won't bring about the good, then it seems that I can rationally intend and act on commitments that undermine the good, without undermining my commitment to it. And we can't appeal to the structure of the commitment to

the good to override the inefficacy problem, because what is at issue is why this particular commitment should come out ahead. Because of the inefficacy problem, these commitments are not strictly conflicting. Both the good and ostensibly conflicting commitments are co-possible precisely because of the inefficacy problem. There is a possible world in which I have a high-emissions lifestyle, but due to the sacrifices of others, this world is ecologically sustainable. So it seems that the norm of consistency does not deliver a resolution in favor of action. It only makes the problem more vivid.

The norm of instrumental coherence seems more promising. According to the norm of *coherence*, a rational agent intends what they believe to be the necessary means to their intended end (Bratman 2012, 76). As I've argued, if an intended end is a commitment that can only be collectively realized, then the rational agent must intend to participate in the set of actions that together constitute a necessary means to that end, or the agent can't be said to have that commitment. Coherence then presents the following dilemma. On the first horn, if the agent is committed to bringing about the collective good, then she ought to intend to participate in the set of actions that together constitute a necessary means to that end. This will often require personal sacrifice, sometimes to a significant degree. On the second horn – call this the "soft horn", because it is less demanding – if the agent chooses not to intend the necessary means to bring about the collective good, then she has several options.

One is to weaken the strength of her commitment.[13] One can be strongly, moderately, or weakly committed to something. A strongly committed person would indeed be required to undertake the demanding actions required of them, at the expense of self-governance, whereas a less committed person has more slack in their web of reasons. They can do less without

[13] This response may help to address a separate objection that my view is too demanding and would force us to be "moral saints" (Wolf 1982) on pain of irrationality. I thank Leif Wenar for this objection.

compromising their self-governance, simply because they haven't imposed certain rational requirements on themselves by taking on a stronger degree of commitment. But weakening the strength of one's commitment doesn't discharge one from intending at least *some* of actions that constitute part of the necessary means to realizing that commitment. Commitment still requires some degree of action.

Another option is to drop this commitment to the level of a desire, which does not entail adhering to the same rational norms. Nevertheless, a desire gives the agent a prima facie reason to satisfy it (Schroeder 2007), so choosing this "soft horn" of the dilemma does not entirely get her out of intending some of the actions that help to constitute the necessary means to the relevant end. To get out of intending any of the actions that help to constitute the necessary means, the agent should simply drop any desire or commitment to the collective good.

Both these horns of the dilemma allow a rational agent to preserve coherence. Coherence will not tell the agent to plump for one horn or the other. It does, however, make the cost of inaction vivid. That is, inaction does not come for free, licensed by instrumental rationality. It comes at the cost of *being a person who is committed to the bringing about the good*. It is up to the agent to decide for herself whether that is a price worth paying. And if the agent does not choose either one of these horns, but instead retains the desire for or commitment to the good without intending at least some of the actions that constitute the necessary means to that good, then she is practically irrational on that account.

To sum up, I have developed Bratman's planning framework to excavate a reason for action in collective impact situations. Agents who act in these situations, instead of succumbing to the logic of the inefficacy problem, are acting in the service of their own self-governance –

assuming that their commitment to bringing about the good takes priority in their scheme of commitments.

But one might think that plans and commitments change, and that they actually ought to change as we learn more about ourselves and the world (Millgram 2015). The norm of stability of intention thus seems stubborn at best, if not irrational. Why should we stick to our commitments? In particular, why should we stick to our commitment to collective goods, given that we know they are so unlikely to be realized?

I do not have a general answer to the deep theoretical objection about the norm of stability. However, we can again appeal to the specific content of the commitment to the good to buttress the applicability of stability to this particular intention. A better world just *is* something that we shouldn't give up our commitment to. Perhaps this commitment can only be justified intrinsically by the analyticity of the good. But more plausibly, and more in line with the instrumentalist commitments developed in this paper, we should be committed to this good because many of our ordinary desires and commitments cannot be realized if collective harms such as climate change continue apace. The commitment to the good, then, is unlike other personal commitments (such as the commitment to stick with philosophy no matter what), which should be more amenable to revision due to the fact that less hangs on them. We simply may not be able to realize many of our other desires and commitments if we do not alleviate collective harms, so we have an instrumental reason to cooperate in order to redress them.

5. Conclusion

Collective action problems pose a challenge for theories of morality and instrumental rationality due to the inefficacy problem. It seems that any individual can say that they have no

moral reason, to act, given that their action will not make a difference to the overall outcome. Instrumental reasons seem even less promising than moral reasons in this context: it is certainly costly for individuals to act, yet it seems highly improbable that their action will make a difference in the desired way.

I have argued that we should look toward the space of instrumental, diachronic reasons for a reason to cooperate in collective action problems. The Ordinary Person's psychology, which is thought to give rise to the inefficacy problem, contains the reason for action. What I have tried to do in this paper is to show that the Ordinary Person is irrational if she is committed to the good, but fails to cooperate because she allows the inefficacy problem to override her commitments. I have built up an instrumental reason for such a person to act in collective action problems. I argued that collective action problems have a similar structure to intrapersonal, intertemporal commitments, so individual commitment to an outcome that requires collective action provides sufficient reason for cooperation. Then, I argued that the requirements of self-governance put a price on noncooperation. Once we examine the requirements of our commitment to a just and good world, we can find a reason to cooperate within our divided selves.

Let me end with some caveats. I have not attempted to generalize my argument to apply to those individuals without the relevant commitment. The account also applies more strongly to collective *harm* situations rather than collective impact situations in general. Despite these qualifications, I hope that I have at least given an instrumental reason to act to those for whom the inefficacy problem is a true problem. The details of this account will be left to future work. The point, for now, is that inaction in collective impact situations doesn't come for free.

References

Andreou, C. (2006). Environmental damage and the puzzle of the self-torturer. *Philosophy & Public Affairs*, *34*(1), 95–108.

Andreou, C., & White, M. (Eds.). (2010). *The thief of time: Philosophical essays on procrastination*. OUP.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bratman, M. (2003). A desire of one's own. *Journal of Philosophy*, *100*(5), 221–242.

Bratman, M. (2007). Temptation revisited. In *Structures of Agency*. OUP.

Bratman, M. (2012). Time, rationality, and self-governance. *Philosophical Issues*, 22, 73-88.

Bratman, M. (2018). *Planning, time, and self-governance: Essays in practical rationality*. OUP.

Broome, J. (2019). Against denialism. *The Monist*, *102*(1), 110–129.

Budolfson, M. B. (2018). The inefficacy objection to consequentialism and the problem with the expected consequence response. *Philosophical Studies*.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 5–20.

Gilbert, M. (2006). Rationality in collective action. *Philosophy of the social sciences, 36*(1), 3-17.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243-1248.

Kagan, S. (2011). Do I make a difference? *Philosophy & Public Affairs*, *39*(2), 105–141.

Kingston, E., & Sinnott-Armstrong, W. (2018). What's wrong with joyguzzling? *Ethical Theory*

& Moral Practice, *21*(1), 169–186.

Kolodny, N. (2008). The myth of practical consistency. *European Journal of Philosophy*, *16*(3), 366–402.

Kruse, K. M. (2013). *White flight: Atlanta and the making of modern conservatism* (Vol. 89). Princeton University Press.

Brennan, G., & Lomasky, L. (1997). *Democracy and decision: The pure theory of electoral preference*. Cambridge University Press.

Massey, D., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard university press.

Millgram, E. (2015). *The great endarkenment*. OUP.

Nefsky, J. (2011). Consequentialism and the problem of collective harm. *Philosophy & Public Affairs*, *39*(4), 364–395.

Nefsky, J. (2017). How you can help, without making a difference. *Philosophical Studies*, *174*, 2743–2767.

Nefsky, J. (2019). Collective harm and the inefficacy problem. *Philosophy Compass*, *14*(4).

Nguyen, C. T. (2019). Games and the art of agency. *Philosophical Review*, *128*(4), 423–462.

Norcross, A. (2004). Puppies, pigs, and people: Eating meat and marginal cases. *Philosophical perspectives*, *18*, 229-245.

Parfit, D. (1984). *Reasons and persons*. OUP Oxford.

Quinn, W. S. (1990). The puzzle of the self-torturer. *Philosophical studies*, *59*(1), 79-90.

Schroeder, M. (2007). *Slaves of the passions*. Oxford University Press.

Singer, P. (1980). Utilitarianism and vegetarianism. *Philosophy & Public Affairs 9*(4), 325-337.

Sinnott-Armstrong, W. (2005). It's not my fault: Global warming and individual moral obligations. In *Perspectives on climate change: science, economics, and ethics* (Vol. 5).

Elsevier.

Tenenbaum, S., & Raffman, D. (2012). Vague projects and the puzzle of the self-torturer. *Ethics*, *123*(1), 86-112.

Tullock, G. (1971). The paradox of revolution. *Public Choice*, *11*(1), 89-99.