Introduction: Why Commit?

Commitment is quite commonplace and, seemingly, quite significant, since it treats certain options as "off the table." My commitment to teaching my class this morning requires me to close off or put aside the possibility of doing some weight training instead. And my commitment to certain healthy eating practices requires me to close off or put aside the possibility of bringing a box of Twinkies as my lunch. Still, it might seem like commitment is either redundant or irrational – redundant if the option committed to is (taking into account its consequences) preferred over the alternatives, and irrational if the option committed to is dispreferred. But, as will become apparent, there are scenarios in which the ability to commit to a dispreferred alternative is necessary to reap the benefits of cooperation or self-control. This Element focuses on the interaction between cooperation, commitment, and control. Drawing from and building on the existing literature, including my own prior work in this space, I guide the reader through the interesting, challenging, and evolving philosophical terrain where issues regarding cooperation, commitment, and control intersect, adding some new contributions along the way.

As part of illustrating how choices restrained by commitments can sometimes pave the way to better outcomes, Section 1 discusses interpersonal and intrapersonal Prisoner's Dilemma situations and the possibility of a set of unrestrained choices adding up in a way that is problematic relative to the concerns of the choosers involved. The classic Prisoner's Dilemma situation involves two purely self-interested agents who, paradoxically, both do worse when they each follow their own self-interest than they would if they had restrained themselves and cooperated with one another.[1] In light of some key moves and refinements in debates regarding Prisoner's Dilemma situations, which I briefly review, paradigmatic examples of Prisoner's Dilemmas (PDs) now include interpersonal and intrapersonal cases in which the choosers need not be purely self-interested but do have a conflict of ends. In intrapersonal PDs, the choosers are generally assumed to be time slices of a person that (to some extent) discount the utility of future time slices. Like many interpersonal PDs, intrapersonal PDs often take the form of a free-rider problem, in which a good is not realized because each chooser refrains from contributing to its achievement, recognizing their individual contribution as trivial with respect to the realization of the good. Free-rider problems are normally associated with choosers who are partial to

[1] According to Martin Peterson (2015, 1), who corresponded with John F. Nash about the matter, the label "Prisoner's Dilemma" was coined by Albert W. Tucker during a lecture discussing Nash's work. (Tucker was Nash's thesis advisor.) A variation on the sort of example associated with the label is provided in Section 1.

themselves. But, building on debates regarding free-rider problems, I introduce the relatively neglected possibility of what I call "impartially benevolent free riders," understood as free riders motivated by impartial benevolence, and discuss the connection between such free riders and poor self-control. I conclude by emphasizing that, without restraint, certain benefits of cooperation and self-control may be out of reach.

Section 2 focuses on the role of precommitment devices in rational choice. As emphasized in the extensive literature on precommitment devices, such devices can helpfully alter choice situations or incentive structures in a way that makes it impossible or extremely costly to defect from a course of action (or a set of actions) that promises benefits of cooperation or self-control. Given the possibility of precommitting to a course of action, it might seem like Prisoner's Dilemma situations can be straightforwardly solved with precommitment devices once they are recognized and so should not persist. Relatedly, it might seem like an individual or collective that does not use a precommitment device to solve a so-called self-control problem is revealing a gap between proclaimed priorities and actual priorities and so should actually be diagnosed as hypocritical. But, as I've emphasized in my work on procrastination, this abstracts from the fact that suitable precommitment devices might not be easily identifiable or readily available.[2] Careful consideration of the case of the persistent procrastinator reveals how challenges associated with employing precommitment devices can prompt second-order procrastination, even for someone who is genuinely and seriously concerned about her failure to make progress on a proclaimed priority. Section 2 ends with some discussion of a particularly controversial precommitment device that is most commonly considered in discussions of mental illness but is of interest in a variety of cases of temptation. The device in question is a "Ulysses contract" and, as I explain, it seems particularly appropriate in intrapersonal free-rider cases of the sort considered in Section 1.

Section 3 considers the role of resoluteness in rational choice. Resoluteness involves a willingness to stick to a plan even if one believes that what the plan calls for conflicts with the preferences associated with one's current perspective. I survey and critically examine a variety of arguments aimed at supporting the view that resoluteness is a genuine possibility for a rational agent, including arguments that focus on the rationality of non-reconsideration, as well as arguments that first contrast evaluating actions directly and evaluating deliberative procedures and then focus on the rationality of acting in accordance with a beneficial deliberative procedure that allows for counter-preferential choice. Relatedly, I advance some new ideas and reasoning concerning resoluteness and

[2] See, for example, Andreou (2007a).

(what I call) "quasi-resoluteness" via an appeal to the importance of "meta-considerations" and an examination of cases of temptation in which there is a risk of the sort of impartially benevolent free riding introduced in Section 1. The arguments in this section of the Element suggest that incurring costs to precommit to an option by penalizing or eliminating alternatives may be a needlessly roundabout route to solving dilemmas in which certain benefits of cooperation or self-control cannot be realized without restraint. For, given the possibility of rational resoluteness, rational agents can reliably resolve to adhere to a cooperative or self-controlled course and follow through; precommitment devices might then only be needed to compensate for irrational irresoluteness.

Section 4 delves into some relevant complications concerning the nature of actions and the nature of intentions (condensing and synthesizing two strands of my prior work).[3] The complications are tied to the fact that correctly evaluating what an agent is doing at a particular point in time requires recognizing what the agent is doing at that point in time, and this is not a trivial matter, since what an agent is doing *at* a particular point in time is normally not contained *in* that point in time. As such, evaluating what an agent is doing at a particular point in time requires recognizing what dispositions, dynamics, and defaults are in play. After explaining the relevant complications, I explain how they impact some of the previous discussions in this Element. For example, I revisit my earlier discussion of impartially benevolent free riding and suggest that, despite initial appearances to the contrary in the sorts of cases of failure involving free riding motivated by impartial benevolence that I consider, *not* all of the doings (or omissions) occurring at points on the way to failure in relation to the goal at stake are at most trivially detrimental relative to the goal. I also revisit the apparent contrast between evaluating actions directly and evaluating deliberative procedures. I suggest that the contrast may be oversimplified, since one cannot be acting intentionally without integrating certain constraints into one's deliberative framework. I then highlight an alternative contrast, namely the contrast between "on-the-whole" evaluations and evaluations "through time," and suggest that so long as actions and deliberative procedures are evaluated in the same way, either both as wholes or, alternatively, both through time, one will not get conflicting verdicts regarding the best alternative.

Notably, throughout this Element, the notion of "preferences" in play is of preferences as "subjective comparative evaluations" (Hansson & Grüne-Yanoff 2017). Such evaluations can be, but are not necessarily, revealed in choice, since the latter might, in some cases, be explained instead by, for example, habit, an autopilot response, or non-deliberative follow-through on a prior plan.

[3] See, in particular, Andreou (2014a, 2016).

1 Interpersonal and Intrapersonal Prisoner's Dilemma Situations

1.1 Introduction

Prisoner's Dilemma situations illustrate the possibility of a set of unrestrained choices adding up in a way that is problematic relative to the concerns of the choosers involved. Complementing the extensive literature on interpersonal Prisoner's Dilemmas is a growing literature on intrapersonal Prisoner's Dilemmas. After briefly reviewing some key moves and refinements in debates regarding Prisoner's Dilemma situations, which include debates regarding free-rider problems, I introduce the relatively neglected possibility of what I call "impartially benevolent free riders" and discuss the connection between such free riders and cyclic preferences and poor self-control. I conclude by emphasizing that, without restraint, certain benefits of cooperation and self-control may be out of reach.

1.2 The Classic Prisoner's Dilemma Situation

Two criminals have robbed a bank. There is enough evidence to convict each of them of the crime of illegally possessing a firearm, but more evidence is needed to convict either of them of the more serious crime of armed robbery. Although they are partners in crime, the criminals are purely self-interested, and each is looking to get off with as light a sentence as possible. Conviction of illegal possession of a firearm will result in a two-year prison sentence. Conviction of armed robbery will result in an additional ten-year prison sentence (for a total sentence of twelve years). There is a way for each criminal to provide decisive evidence against the other without incriminating himself. Recognizing this, the district attorney offers each criminal the following deal: provide decisive evidence against the other and your sentence will be reduced by a year. Each criminal considers the situation and recognizes that, whatever the other does, his best strategy is to provide the incriminating evidence requested. For if his partner provides incriminating evidence, he does better by providing incriminating evidence too, since doing so allows him to reduce the sentence he will face from twelve years to eleven. And if his partner does not provide incriminating evidence, he still does better by providing incriminating evidence, since doing so allows him to reduce the sentence he will face from two years to one. Since both criminals are purely self-interested, they each provide the district attorney with incriminating evidence and each ends up with an eleven-year sentence. Of course, had they both refused to provide any incriminating evidence, they would have each ended up with a two-year sentence. The district attorney's

offer has thus landed them in a dilemma: self-interest prompts each of them to incriminate the other, but in both following their self-interest, they both end up worse off than they would have ended up if they had both restrained themselves.[4]

1.3 Prisoner's Dilemmas among Selfless Individuals

This Prisoner's Dilemma captures the essential feature of choice situations that now go by the name of Prisoner's Dilemmas. Roughly put, in a Prisoner's Dilemma, choices prompted by each chooser's concerns add up to an outcome that is worse, from the perspective of each chooser, than another that was available.[5] No prisoners need be involved; more than two choosers can be involved; and, interestingly, the choosers need not be purely self-interested. Indeed, a Prisoner's Dilemma situation might obtain even among totally selfless individuals if their selfless ends do not coincide. Consider, for example, Peter Vanderschraaf's shipwreck case, in which all that remains from the provisions that were brought along is one biscuit:

> If the survivors of a shipwreck are all perfectly selfless angels, then each will be tempted to leave all of what little food they have for the others. But if each angel yields to temptation completely, all will starve, even though each angel is acting so as to prevent the others from starving. Knowing this, the angels need to divide the food so that all get shares, even though this requires each of them to yield somewhat to the wishes of the others. (2006, 330)

Assuming that each selfless angel favors leaving the whole biscuit for others over eating some of it (knowing that all she can control is how much she leaves for them, not whether they will eat it), but also prefers that everyone eats some of the biscuit than that none eat any of it, this case involves a dilemma of the same structure as that in the original PD case presented: If each chooser proceeds selflessly (by, in this case, refusing to eat any of the biscuit), all do worse relative to their aim (of benefiting the others) than if they all show restraint relative to their aim (and all eat a bit themselves). That each agent's end is selfless does not interfere with the possibility of a PD being generated.

[4] As indicated in footnote 1, this is a variation on the sort of example associated with the label "Prisoner's Dilemma."

[5] More precisely, in a Prisoner's Dilemma, (1) each chooser has a dominant strategy in the sense that there is a single choice that invariably serves as her best possible choice, given her concerns, no matter how the others involved choose, and yet, (2) if each chooser follows her dominant strategy, the result is an outcome that is dominated in the sense that there is another possible outcome that all the choosers involved would have preferred. Notably, since each chooser in a PD has a dominant strategy, knowing how the other(s) will choose does not provide her with relevant information regarding her best choice.

1.4 Prisoner's Dilemmas, Cooperation, and the Circumstances of Justice

The preceding example figures in discussions regarding the connection between cooperation and justice. The circumstances of justice are prominently construed as the conditions under which a willingness to cooperate, where this involves a willingness to show restraint, or be restrained, for the sake of mutual benefit, is rationally called for. Thomas Hobbes famously cast justice as the rational solution to the chaos and destruction that purely self-interested agents – who treat others as a mere means to their own "conservation" and "delectation" – collectively generate in the absence of enforced ground rules (1668/1994, 75). David Hume (1951/1998, section 3, part 1) affirmed a connection between self-interest and justice when he argued that "extensive benevolence" would make justice redundant, since, insofar as one cares for others, one will take their interests into account even if there are no demands of justice compelling one to do so. But, for Hume, our need for justice is tied to our *limited* self-interest, which does not involve a complete indifference toward others but does involve "more concern for [our] own interest than for that of [our] fellows." Later theorists, including, perhaps most influentially, John Rawls, loosened the connection between self-interest and the circumstances of justice even further and sided with the now prevailing view that the circumstances of justice can obtain even among individuals who are not self-interested if they have different "plans of life" and "conceptions of the good" (Rawls 1971, 127) or, indeed, any conflict of ends, even if the ends they are committed to are not at all selfish. Relatedly, paradigmatic examples of PDs now include cases in which the choosers are not purely self-interested but have a conflict of ends.

1.5 Intrapersonal Prisoner's Dilemmas and Discounting

The recognition that PDs can be generated even among choosers who are strongly connected to one another by ties of care and concern facilitates the recognition of the possibility of *intra*personal PDs. Unlike in interpersonal PDs, in intrapersonal PDs, the choosers at issue are not separate individuals but the same individual at different times. Such PDs are easily generated when someone cares about her future (self) but also discounts future benefits, especially distant future benefits, assigning more weight to benefits that are imminent and so loom large.

It is generally recognized that human beings discount future utility, often favoring smaller, sooner rewards over larger, delayed rewards (e.g., \$10 of treat-money today over \$11 of treat-money tomorrow). Agents who discount

future utility are fragmented into (what are sometimes referred to as) time-slice selves. Each time-slice self is not indifferent to the fate of the other time-slice selves, but closer time-slice selves are favored over more distant time-slice selves. Intrapersonal PDs exist when each time-slice self favors the achievement of a long-term goal but also prefers that the restraint needed to achieve the long-term goal be exercised not by her current self but by her future selves.

Suppose, for example, that I want my far-future self to experience a comfortable retirement and that this will require me to frequently save small amounts while I am still young. I have a little money that I can save today and put toward my retirement funds, but I consider the possibility of all (or most of) my future selves (with extra money) saving a little and realizing my long-term goal without any contribution from my current self, and I prefer that instead. Moreover, if I expect that none or few of my future selves (with extra money) are going to save, my long-term goal will not be met regardless of whether my current self contributes, and so my current self still favors contributing nothing. Still, all my time-slice selves prefer that we all save a little toward a comfortable retirement than that none do. I thus face an intrapersonal Prisoner's Dilemma situation.

It might be objected that if I really do discount future utility, favoring more proximate future time-slice selves over more distant future time-slice selves, I will not want the former time-slice selves to make any sacrifices for the latter time-slice selves, and so it will not be true that all of my time-slice selves prefer that they all show restraint than that none do. Notice, however, that discounting future utility need not involve invariably favoring rewards for earlier time-slice selves (no matter how minor) over rewards for later time-slice selves (no matter how crucial); depending on how exactly, and how much, future rewards are discounted, there is room for discounters to favor the sacrifice of small luxuries by more proximate future selves to ensure funds for necessities for more distant future selves.

But, if one discounts future utility at a rate that prompts one to currently favor spending a small amount of discretionary money today over saving it for twenty years from now, won't one also currently favor spending a small amount of discretionary money when one gets one's next paycheck (in the not-too-far future) over saving it for twenty years from then? The crucial thing to notice here is that while the answer would be "yes" given a fixed discount rate of, say, *n* percent per day, one's discount curve need not reflect a steady discount rate. And if (as evidence suggests is the norm) one's discount curve does not reflect a steady discount rate, it can be such that, while one currently favors spending a small amount of discretionary money today over saving it

for twenty years from now, one also currently favors showing restraint when one gets one's next paycheck and saving a small amount of discretionary money then for twenty years from then.[6]

1.6 Interpersonal and Intrapersonal Free Riding Motivated by Partiality

The preceding intrapersonal Prisoner's Dilemma situation takes the form of a free-rider problem. In classic interpersonal free-rider problems, an agent who is partial to himself seeks to obtain a certain benefit without contributing to the realization of the benefit himself. Suppose, for example, that one wants to enjoy fairly clean air and that this requires a major reduction in the amount of polluting emissions in one's area. There is a major campaign encouraging individuals to take the bus to work rather than drive. One hopes the campaign will result in a significant improvement in air quality but, recognizing that whether this will occur does not hang on whether one participates because one's own participation would have no more than a trivial impact, one opts to proceed as usual and, hopefully, free ride off the efforts of others. The problem is that, insofar as others are similarly motivated, the result is "a tragedy of the commons." In a tragedy of the commons, a valued communal resource, such as clean air, is lost because, even though everyone would rather that all show restraint, with the result that the resource is preserved, than that none show restraint, with the result that the resource is lost, each acts on the recognition that, whatever others do, she is better off not showing restraint.[7]

In the retirement case, each time-slice self, though not indifferent to her future selves, is, nonetheless, somewhat partial to her current self and so favors shifting the burden of achieving the valued outcome of a comfortable retirement onto her future selves. In this case, tragedy results when the burden is not taken on by any of the time-slice selves, and the far-future selves that none of the time slices wanted to see suffer are miserably destitute.

1.7 The Impartially Benevolent Free Rider

Perhaps the most interesting under-explored aspect of free-rider situations is that they seem to make room for Prisoner's Dilemma situations in which the agents involved are all impartially benevolent.[8] (Of course, this possibility is precluded if it is *stipulated* that PDs must involve choosers who do not

[6] See Ainslie (2001, chapter 3).

[7] The label "tragedy of the commons" can be traced to Hardin (1968). For an influential general discussion of the logic of collective action, see Olson (1965).

[8] See Andreou (2010, 207) for some remarks that verge on recognizing this possibility. This aspect of free-rider situations is related to the puzzling "moral mathematics" in cases like Derek Parfit's

fully share ends; but this stipulation can appropriately be put aside if the dilemma in PDs – which is, roughly put, that choices prompted by each chooser's concerns add up to an outcome that is worse, from the perspective of each chooser, than another that was available – is possible even for choosers who do fully share ends.) Consider the following case. J wants to lose enough weight so that he looks about as trim as he was when he got married (and fit into size 32 pants). It is easy to see how J might find himself in a dilemma if he discounts future utility. But suppose he does not discount future utility; suppose, in particular, that his time slices are not at all partial to themselves but are impartially benevolent. A dilemma can still arise. For, as long as it is true that whether the desired outcome will occur does not hang on whether this particular time slice, say J_n, sacrifices his enjoyment by, for instance, passing up a brownie bite (since J_n's sacrifice will have no more than a trivial impact), J_n, as well as his past and future time slices, can, motivated by impartial benevolence, reason as follows: J_n's sacrifice will have no more than a trivial impact on J's appearance; so, whatever his future time slices do, it is preferable that J_n enjoy the brownie bite and put off dieting for now. Even an impartially benevolent agent can favor allowing for some free rides when the cost to the whole is trivial enough. The problem, of course, is that there is nothing distinctive about J_n, and so the sort of benevolent thinking under consideration can, like free riding motivated by partiality, result in an important goal never being met.

The same sort of challenge can cause trouble for a unified collective. Consider, for instance,

> a collective that values a healthy community, values luxuries whose production or use promotes a carcinogenic environment, and believes that if it does not curb its consumption, the health of the community will be seriously damaged. This collective might rightly think that if it is going to curb consumption, it is better off starting next month rather than right away. For the community can then enjoy another month of luxury living and this will not take the community's members from a state of decent health to a state of poor health. Yet, if the collective opts for a high level of consumption month after month, the health of the community will be seriously damaged. (Andreou 2006a, 104)

1.8 Unification, Fragmentation, and Cyclic Preferences

Though they can be divided into component time slices or selves, both the temporally extended individual J and the collective just described are *unified*

"Drops of Water" case (1984, 76), at least if contributing one's pint of water has some cost (of, e.g., effort or opportunity). See, relatedly, Glover (1975), Kagan (2011), and Temkin (2012).

in the sense that, by hypothesis, there is no conflict of ends *between* different choosing slices or selves; all the choosing slices or selves (including every time slice, J_x, of J in J's case and, alternatively, every time slice of the collective in the community case) have the same shared interests. Each choosing slice or self is, however, *fragmented* in the sense that it itself has potentially conflicting interests that it must consider when it is making choices. For example, J_n must consider both his interest in J's appearance and his interest in J's culinary delight – interests that he shares with all of J's time slices. Such fragmentation can raise the same sorts of challenges to effective decision-making as is faced by agents with conflicting ends. Notice, in particular, that such fragmentation can generate not only a Prisoner's Dilemma situation but also related *cyclic* preferences.

Preferences over a set of options qualify as cyclic if the options cannot be ordered from most preferred to least preferred (even allowing for ties) because the preference structure puts the options in a "loop" wherein, for example, O_1 is preferred to O_2, O_2 is preferred to O_3, . . ., O_{n-1} is preferred to O_n, but O_n is preferred to O_1. We know from Condorcet's paradox that if a group of agents who do not share preferences vote on options in accordance with their preferences, the resulting "social preference" can be cyclic, even if each agent's preferences are not.[9] For example, if A's ordering of O_1, O_2, and O_3 from most preferred to least preferred is O_3, O_2, O_1, B's ordering is O_2, O_1, O_3, and C's ordering is O_1, O_3, O_2, then the "social preference" that results from majority rule pair-wise voting on the options results in a preference loop in which O_3 is preferred to O_2, O_2 is preferred to O_1, but O_1 is preferred to O_3. Relatedly, an agent with stable preferences, but *multiple* interests, and a prioritization system that appeals to thresholds, can have cyclic preferences even if the agent's ranking of a set of options relative to a single dimension of concern is invariably acyclic (i.e., not cyclic). Suppose, for example, that O_n is the option of consuming *n* brownie bites this month, and that, in terms of their fit with culinary delight, A's ordering of options O_1, O_2, . . ., O_{300} from most preferred to least preferred is O_{300}, . . ., O_2, O_1; by contrast, in terms of their fit with dieting, A's ordering of options O_1, O_2, . . ., O_{300} from most preferred to least preferred is O_1, O_2, . . ., O_{300}. Suppose further that, taking into account both factors, A prefers an option that fits with dieting above an option that fits with culinary delight unless the difference between the options in terms of their contribution to weight loss is too small to make a significantly noticeable difference to anyone judging A's appearance. Then, as suggested by J's case,

[9] Condorcet's paradox is named after the Marquis de Condorcet, who pinpointed the paradox as part of his pioneering work in social choice theory. See de Condorcet (1785).