Putting Willpower into Decision Theory:
The person as a team over time and intra-personal team reasoning

Natalie Gold

Senior Research Fellow, Department of Philosophy, University of Oxford

Principal Behavioural Insights Advisor, Public Health England

for J. Bermudez (ed.) *Self Control and Rationality* Cambridge: Cambridge University Press.

Abstract. In decision-theory, problems of self-control can be modelled as problems of intra-personal cooperation, between a series of transient agents who each make choices at particular times. Early agents in the series can try to influence the actions of later agents, but there is no rational way to exert willpower. I show how willpower can be introduced into decision theory by applying the theory of team reasoning, which was originally developed to understand cooperation between individuals in groups and allows that there can be multiple levels of agency, the individual and the team. In the case of intertemporal choice, the levels are the transient agent and the person over time. Intra-personal team reasoning, understood as a psychological process of identifying with the person over time, can generate a plausible theory of rational control if the intertemporal problem is structured as a threshold public goods game. In this framework, willpower is the ability to align one's present self with one's extended interests by identifying with the person over time. I show how intra-personal team reasoning creates a space for resolutions in decision theory and how it resolves a puzzle that exists in accounts that understand willpower as making and then not reconsidering resolutions.

Keywords: Agency, Decision theory, Framing, Intentions, Intertemporal choice, Intra-personal cooperation, Resolutions, Self-control, Team reasoning, Willpower

1. Introduction

In decision theory, a standard way of modelling a person who has to make a series of choices over time is as a series of transient agents, who each make choices at particular times. In this framework, a person has a problem of self-control when the course of action that is preferred by an earlier transient agent relies on a choice by a later transient agent that will not be preferred from that later transient agent's point of view. Faced with such a problem, the earlier transient agent can try to influence the later transient agent's behaviour by taking action to alter her later self's incentives or limit her later self's opportunities. However, what she cannot do is use willpower. A person who believes that she can simply make a plan and act on it when the time comes is considered "naive" and her naivety is a cause of bad outcomes. This view is in stark contrast to much research in psychology and philosophy, which assumes that people have willpower, and that exercising willpower can be rational and lead to good outcomes.

The model also has a lacuna when it comes to agency. Problems of self-control are caused by conflicts between transient agents; there is no sense of an agent who has interests that extend over time. The transient agents may care about the outcomes of the other transient agents in the series, which may figure in their preference functions. Nevertheless each transient agent acts on its own transient interests. Related to this, the standard model lacks the concept of an intention that guides behaviour over time. To the extent that the model includes intentions, they are merely predictions of future behaviour that turn out to be correct. This connects back to self-control because some philosophers have argued that willpower consists in not revising one's intentions (Holton, 1999).

I will show how willpower can be introduced into decision theory—and the gap between psychology, philosophy and economics bridged—by allowing that there can be multiple levels of agency and applying the theory of team reasoning. Team reasoning is an extension of game theory which allows that there can be multiple levels of agency (Sugden, 1993, Bacharach, 2006). In standard game theory, the only recognised level of agency is the individual, analogous to way that models of the person over time only allow one level of agency, the transient agent. Team reasoning was developed to understand the behaviour of individuals in groups. The idea is that not just individuals but also groups

can be agents, so that rather than asking "What should *I* do?", the individuals in the group can ask "What should *we* do?", which can solve problems of coordination and cooperation. Team reasoning extends game theory to allow more than one level of agency. Although the levels of agency in the original applications are the individual and the group, the conceptual apparatus can be applied to other problems. Thus, we can use the theory of team reasoning to introduce a second level of agency, the person over over time, into the model of intertemporal choice, as well as the transient agents.

Intra-personal team reasoning can explain why it is rational to exert self-control without abandoning the decision theoretic framework and it can provide a basis for incorporating intentions into game theory. After introducing the intertemporal problem, I will highlight some parallels with inter-personal problems of coordination and cooperation, in order to motivate the use of team reasoning. Then I will introduce team reasoning in the inter-personal context, before showing how it can be applied to the intra-personal case.

2. Decision Theoretic Models of the Problem of Self-Control

In decision theory, problems of intertemporal choice are often analyzed as if, at each time *t* at which the person has to make a decision, that decision is made by a distinct transient agent or "timeslice", the person at time *t*. Each timeslice is treated as an independent rational decision-maker, so that "the individual over time is an infinity of individuals" (Strotz, 1955, p.179). This does not imply any metaphysical commitments, in particular it is not an endorsement of perdurantism, the view that things really do consist of temporal parts. Rather, it is a natural way of modelling people because the self at a particular time is the locus of choices, experiences, and perceptions.

Choices are events that are located in time. Choices result in experiences and an experience is also a type of event, had by a person at a time or over a temporal interval. An experience can be pre-experienced or re-lived, through anticipation and memory; anticipation and memory are also types of experiences, which occur at a time, even though they are one step removed from the initial stimulus.

First person perception of events is also had by the self at a time. There is some evidence from psychology that, when we contemplate our experiences, as we get further away from the present we tend to take a third person perspective. When people are asked to imagine a scene from their past or their future, they are more likely to see themselves in the picture, rather than seeing the scene as though through their own eyes (Pronin and Ross, 2006). Nor are we good at predicting our future preferences and attitudes, or what our future experiences will feel like (Lowenthal & Loewenstein, G., 2001; Loewenstein, O'Donaghue, & Rabin, 2003; Van Boven, & Loewenstein, 2005). There is a gap between our knowledge of our current experience and our future.

The timeslice model captures the idea that choices are made at a time, by selves that have a first person perspective on that time and a third person perspective on the past and future.

We can think of the successive selves as involved in strategic interaction or the self as a community of interests (Schelling, 1984; Ainslie, 1992), which gives a nice framework in which to set up problems of self-control in the face of temptations. Here is an example from a classic paper by O'Donaghue and Rabin (1999). Suppose you usually go to the movies on Saturday nights. The schedule for the next four weeks is as follows: week 1 is a mediocre movie, week 2 a good movie, week 3 a great movie, and week 4 a Johnny Depp movie, which is best of all. You also have a report to write for work, due within a month, and in order to write it you know that you will have to stay in one Saturday night and must therefore skip a movie. The question is: when do you complete the report?

It seems obvious that the best overall plan is to do the report on the first Saturday. That is the option that would be chosen by a planner who was working out your schedule in week 0. But all that we need to add is a little bit of present bias, with the current timeslice favouring itself, for you to miss the Johnny Depp movie. In order to see this, let us suppose, following O'Donaghue and Rabin (1999), that the valuation of the mediocre, good, great, and Depp movies are 3, 5, 8, and 13 and that the cost of writing the report is just the cost of not seeing the movie that evening. It is plausible that each timeslice gives more weight to its own experiences. Imagine that each timeslice places double the weight on its own experiences than those of other timeslices. In that case, a naive agent, who

believes that she can make a plan and act on it when the time comes—even in the face of temptation—will end up missing the Johnny Depp movie. Come the first Saturday night the timeslice, call it T1, based on her current valuation, judges that she should go to the mediocre movie (which, with double weight, is valued at 6) and skip the good movie next week (valued at 5). But, next week, T2 finds herself in exactly the same situation. She justifies to herself why a night out is particularly valuable to her right now, so she chooses to go to the good movie tonight (now valued at 10), believing she has the willpower to skip the great movie next week (currently valued at 8). The same happens with T3, leading to the situation where the agent is forced to miss the Johnny Depp movie in week 4.

There are three things to note about this example. First, the timeslice that writes the report bears a cost (missing the movie that week) for which others get the benefit (they get to go to the movie other weeks). In economic language, there is an "externality" because the agent's choice has consequences that affect other agents. Second, each timeslice magnifies the sacrifice that would be made by herself but does not realise that other timeslices will also have a "present-bias". However, naivety about future selves is not essential to the problem. Even a "sophisticated" agent, who has correct expectations about her future present bias and backward inducts accordingly, will procrastinate for a week in Rabin and O'Donaghue's model.[1] Third, the timeslices end up with a outcome that is ranked very low by all timeslices. Missing the Depp movie is everyone but T3's worst outcome, and it is T3's second worst outcome after writing the report herself and missing the great movie. Conversely, apart from T1 who prefers that the report is written in week 2, all the timeslices prefer writing the report in week 1.

---

[1] Reasoning by backwards induction: There will be no choice in week 4. If she has not written the report then she will have to skip the movie. She can also predict that, if she gets to week 3 and has not written the report, then she will end up missing the Depp movie. Given that, if she gets to week 2 without having written the report, then her effective choice would be between missing the good movie and missing the Depp movie. According to the model, that's a big enough difference in payoff that the T2 would prefer to skip the good movie in week 2 than have the T4 miss the Depp movie. So if she has not written the report by week 2, she will write it in that week. Thus, in week 1, the choice is between skipping the mediocre movie in week 1 and skipping the good movie in week 2, and the T1 prefers to skip the good movie, so she does not write the report in week 1. Come week 2, T2 writes the report.

Decision theory offers two types of resources for an agent who, in week 0, wants to ensure that she will overcome temptation in week 1. The T0 self can change the incentives faced by her future selves or she can alter her opportunities (Thaler and Sheffrin, 1981; Greer and Levine, 2006). For example, she could enter into a side-bet with a colleague, which entails paying out a large sum if the report is not done in the first week, or she could she could rip up her week-1 cinema ticket.

These are intended as examples of changing incentives versus destroying options, but there is a very fine line between the two strategies. Many pre-commitments, which destroy options, may actually be a way of making an option more costly: when one destroys an option there is usually the possibility of replacing it, albeit at some cost. In the previous example, the cinema ticket could presumably be replaced; even if the movie is sold out, there would be some price at which another ticket holder would trade.

In decision theory, a person who has a problem of self-control in the face of temptation faces a foreseen preference change. In the model, if we observe someone who does not give in to temptation, then either an earlier timeslice took action to change the incentives, or else she never had conflicting preferences in the first place.

3. The strange lack of self over time

Many philosophers and psychologists believe that we have another resource for resisting temptation: willpower. A popular idea is that willpower involves making "resolutions", or plans whose purpose is to help us overcome temptation (Holton, 1999, 2009). Decision theory has some room for plans—if they are incentive compatible i.e. if all the stipulated choices will maximize utility at the time of choice and therefore a decision-maker can predict in advance that she will make them. This reduces planning to prediction. I will show how decision theory can make room for willpower and for plans that guide action. Although my solution includes an explanation of intentions and resolutions, the basic capacity for self-control is intra-personal team reasoning. Therefore my solution explains willpower, but in a different manner from Holton's. In fact, the dependency is in the other direction: we cannot make sense of resolutions within decision theory without adding something like my proposed mechanism of willpower.

I motivate my approach by noting an oddity of the standard decision theoretic model of the agent over time, namely the strange lack of the self over time. The transient agents in the O'Donaghue and Rabin (1999) model put weight on the outcomes of the other timeslices. Never-the-less, the preferences in the model are those of the transient timeslices, there is no sense in which they can hold preferences qua continuing self over time. Yet we usually think that people have interests that extend over time.

This way of modelling people, without extended interests or extended selves, has some counter-intuitive implications. We can see these very clearly using a simple example where the transient agents share a common goal. Take someone who wants to cross the road, from east to west on a two-lane city street. In order to cross the street, she must perform two actions in sequence. In period 1 she must walk from the east side of the street to the middle. Then, in period 2, she must walk from the middle to the west side. From the perspective of conventional decision theory, there are two transient agents, the agent in period 1 (T1) and the agent in period 2 (T2). Rational agents reason by backwards induction. So, in period 1, the agent's reasoning (as T1) will go something like this: I can either stay on the east side or go to the middle. If I go to the middle, then T2 will then have to choose whether to go on to the west side, or to return to the east side. Since I expect T2 to want to be on the west side, I deduce that she would go on rather than back. So, since I want T2 to get to the west side, I should go to the middle. T1 accordingly crosses to the middle of the road. Then, in period 2, T2 notes that she is in the middle of the street, and reasons: I would rather be on the west side than the east, so I should go to the west side.

This type of reasoning gets the person to the other side of the street. However, it feels intuitively odd: the "I" in the reasoning refers to the transient timeslice, not the person over time. There is an absence of any sense of agency over the whole period, of a continuing self who has interests that extend over the whole period and who can form an intention to cross the street and then just carry it out. In period 1 our agent can predict that she will continue the action in period 2. However, in neither period can she think of herself as performing the action of crossing the street; she cannot perceive herself as a continuing agent, nor act on her intentions qua continuing agent.

A standard decision theorist might counter that she would not usually model crossing the street in this manner. Decision theory is flexible about the length of time for which a transient agent exists. It is not committed to the fleeting timeslices found in metaphysics. One would not usually model separate transient agents when the transient agents' interests are aligned because this situation does not usually present an interesting problem. There are two replies to this.

The first is that the timeslice modelling strategy is not merely confined to situations of conflict of interests, people *have* modelled transient agents whose interests are aligned. Even when interests are aligned, there are interesting and perplexing issues. In particular, decision theory cannot necessarily predict that transient agents will coordinate over time in sequential coordination games (Binmore, 1987; Pettit and Sugden, 1989; Reny, 1992; Gold and Sugden, 2006).

The second response is that the decision theorist is still missing something interesting. The intra-personal coordination example was supposed to show how the model does violence to our intuitions by not acknowledging agency over time. However, it is not critical to view intra-personal coordination as a problem in order to think we need to supplement the standard model in the case of self-control. For the decision theorist, there is a problem of self-control when interests conflict, but no problem when they are aligned. The standard phenomenology of temptation involves feeling conflicted between long- and short-term interests, and a natural way to think of self-control is the ability to align one's short- and long-term interests. Decision theory does not include this conflict, nor does it explain how some agents can resolve it without the use of external crutches. Standard decision theory has nothing to say about why preferences are sometimes aligned and other times they are not.

Decision theory provides a neat model of lack of self-control, but has a lacuna when it comes to self-control. The crucial feature in the decision theoretic model of self-control is a temporary and anticipated preference change. In order to exhibit self-control, an agent must bring her future preferences into line. In the model, this can only be done by changing the environment. What the model lacks is an internal mechanism by which an agent can bring her preferences into line, or an explanation why, absent external

mechanisms, one agent can exercise self-control in the face of temptation when another agent cannot.

4. Inter-personal Team Reasoning

While the transient agent model does very well at capturing the conflict of interests between timeslices that can lead to problems of self-control, it does not capture our sense of agency over time or the role of willpower, in the sense described in philosophy and psychology. In order to capture agency over time and willpower without losing the insights of the transient agent model, we can introduce another level of agency, the person over time, as well as the level of the transient agents. In other words, we need a model of multiple levels of agency. Luckily, such a model already exists in the inter-personal case, which we can apply to our problem.

The theory of *team reasoning* was motivated by two families of game that have counter-intuitive solutions (Colman & Gold, 2017, Karpus & Gold, 2017). One family of games has multiple Nash equilibria, but one of the equilibria *Pareto dominates* the others—there is one outcome in which all players are better off—yet game theory cannot recommend or predict the strategies that lead to the Pareto-dominant outcome. All classical game theory can say is that the rational solution will be one of the Nash equilibria. One member of this family is the coordination game known as Hi-Lo, shown in Figure 1, another is the Stag Hunt. We can illustrate the problem using the Hi-Lo game. Standard game theory says that a rational player will choose a best reply to the other player(s). If P1 chooses *high*, then the best response by P2 is also to choose *high*, so (*high*, *high*) is a Nash equilibrium. However, if P1 were to choose *low*, then the best responses by P2 is to choose *low* as well; if P1 chooses *low* and P2 chooses *high* then both get nothing. Therefore (*low, low*) is also a Nash equilibrium. Standard game theory recommends that rational players will play their parts in a Nash equilibrium, but it cannot advise one Nash equilibrium over another, so it cannot recommend to the players that they both play *high.*

P2

		high	*low*
P1	*high*	2, 2	0, 0
	low	0, 0	1, 1

Figure 1: Hi-Lo game

The second family of games is those with a single Nash equilibrium that is Pareto dominated by a non-equilibrium outcome. In this case, game theory would recommend and predict that the strategies leading to the non-equilibrium outcome will *not* be played. An example of this type of game is the infamous prisoner's dilemma, or its multi-player version, the public goods game.

Team reasoning can explain why it is rational for individual players to choose the strategies that lead to the Pareto-dominant outcomes. The idea is that, when an individual identifies with and reasons as a member of a team, she considers which *combination* of actions by members of the team would best promote the team's objective and then performs her part of that combination. Instead of asking "What should *I* do?" as per classical game theory, players can ask, "What should *we* do and how can I play my part?". It is clear that, if there is common knowledge that all players group identify and are team reasoning, the theory of team reasoning can recommend and predict *high*-play in Hi-Lo. In the prisoner's dilemma, if the off-diagonal outcomes are viewed as worse than the (*C*, *C*) outcomes from the perspective of the team, then with common knowledge of team reasoning the theory can predict and recommend *C*-play. (For a more detailed explanation see Gold & Sugden, 2007a, Gold & Sugden, 2007b.)

Team reasoning involves both a payoff-transformation, to what Sugden (1993) calls "team-directed preferences", and an agency transformation, taking the relevant unit of agency to be the group. In behavioural economics, theorists often start with the material payoffs that subjects face and talk of their transformation into the utility payoffs that guide behaviour, which may diverge from their material payoffs (for instance if they care about what the other player gets). In the theory of team reasoning, we start with the utility payoffs that represent what the player wants to achieve as an individual and, when an

individual group identifies, the payoff transformation is to the payoffs that the player wants to achieve as a team member (Gold, 2012). Payoff transformation alone will not suffice, the agency transformation is a necessary part of the process. To see why, consider what would happen if we only had payoff transformation. No plausible payoff-transformation will change the ordering of the payoffs in Hi-Lo, where interests are already aligned. (See Karpus & Gold, 2017 or Colman & Gold, 2017 for a more extended explanation.) In the prisoner's dilemma, payoff transformation theories usually turn the (*C, C*) outcome into an equilibrium, but they do not change the equilibrium status of the (*D, D*) outcome, so they still cannot predict cooperative choices (Gold, 2012, provides more detail). In order to see this, take the prisoner's dilemma on the left-hand side of Figure 2. Transforming the game using golden mean altruism, where each player is motivated to maximize the average of the player's outcomes, gives the matrix on the right-hand side of Figure 2, which is a Hi-Lo. (See also Gold & Sugden, 2007a, Gold & Sugden 2007b.)

		P2				P2	
		C	*D*			*C*	*D*
P1	*C*	4, 4	0, 5	P1	*C*	4, 4	2.5, 2.5
	D	5, 0	3, 3		*D*	2.5, 2.5	3, 3

Figure 2: Prisoner's Dilemma and Prisoner's Dilemma transformed

Team reasoning was developed separately by Bacharach (1997, 1999, 2006) and Sugden (1993, 2000, 2003), and they have different explanations about when and why team reasoning occurs. Both Bacharach and Sugden's theories involve framing and expectations, but Bacharach's emphasis is on framing while Sugden's is on expectations.

For Bacharach, team reasoning is a psychological process. Whether or not someone team reasons simply depends on whether she "frames" the game as a problem for "me" or a problem for "us". In an *unreliable team interaction*, there is some doubt as to whether other team members group identify and team reason. When deciding what to do, someone who

team reasons will use *circumspect team reasoning*, taking into account the probability that other players team reason and maximizing expected utility from the perspective of the team. For instance, in the prisoner's dilemma, cooperating may not maximize the team utility if there is a large enough chance that other player does not team reason, so circumspect team reasoning does not lead to unconditional cooperation. However, for Bacharach, team reasoning does not follow from rationally accessible deliberation and team reasoning may leave the individual worse off in terms of her individual lights, for instance if the team were to rank the off-diagonal (*C, D*) and (*D, C*) outcomes of the prisoner's dilemma higher than the (*C, C*) outcome, or if circumspect team reasoning recommends that *C*-play would maximize expected utility ex ante, but ex poste the other player turns out not to have group identified.

For Sugden, team reasoning is a part of a social contract theory, where an individual can choose to cooperate with others for mutual advantage (Sugden, 2011, 2015). If an individual sees that it is possible to frame a game as a problem for "us", then she may decide to team reason. However, no individual would team reason unless it furthered her individual interests, which puts constraints on the team payoff ordering. Sugden's team reasoners will not risk getting suckered, which also means that they will not team reason without assurance that others are team reasoning too, hence we can call it *mutually assured team reasoning*. The idea that people can decide to team reason, the constraints on the team preferences, and the need for assurance are all points of difference with Bacharach's theory. (See Gold, 2012, for more detailed comparison.)

5. Intra-personal team reasoning

Problems of self-control are problems of intra-personal cooperation (Gold, 2013). The classic inter-personal problem of cooperation is the prisoner's dilemma, a type of *public goods game*, where costly actions by individuals have positive externalities, so that it is individually rational for an individual to defect (not to contribute to the public good), even though all individuals prefer the situation where everyone contributes to the one where no-one contributes. In other words, one agent takes an action whose benefits are spread across many agents. The benefit that accrues to the individual does not outweigh

the cost of the action, but the benefit that accrues across all individuals does. Problems of self-control are similar in that they involve one transient agent paying a cost in return for benefits that accrue to other transient agents, so they are problems of intra-personal cooperation.

By analogy, if inter-personal team reasoning can lead to cooperation in the prisoner's dilemma, then intra-personal team reasoning can promote self-control. In the intra-personal case, the team consists of the set of timeslices that make up the person over time. The units of agency are the timeslice and the self-over time, and the equivalent of group identification is identifying with the person over time. In the standard model, the timeslice does "transient-agent reasoning", asking "What should *I*-now do?", whereas intra-personal team reasoning allows any timeslice that identifies with the person over time to ask "What should the *I*-the person over time do?" and to play its part in the best team plan.

Take the problem of Jo's examination, introduced by Gold and Sugden (2006). This is a three period model, with three transient agents, Jo_1, Jo_2, and Jo_3. In period 3, Jo takes an examination, in periods 1 and 2, Jo must decide whether to study for the exam or to relax. The experienced utility (in the sense of Kahneman and Thaler, 2006) of studying in any period is –3, while that of resting is 0. In period 3, experienced utility is 0 if Jo has rested on both previous days, 5 if she has rested on one day and studied on the other, and 10 if she has studied on both. In terms of experienced utility, if either Jo_1 or Jo_2 chooses to study, then that timeslice bears a cost that has benefits for Jo_3. The benefits of studying are greater over the lifetime than the costs. However, the transient agents that study do not capture the benefits. Figure 3 shows the payoffs for each combination of moves in terms of experienced utility.

		Jo_2	
		study	*rest*
Jo_1	*study*	–3, –3, 10	–3, 0, 5
	rest	–3, 0, 5	0, 0, 0

Figure 3: Jo's examination, experienced utility payoffs for Jo_1, Jo_2 and Jo_3

		Jo_2	
		study	*rest*
Jo_1	*study*	1, 1, 14	–1, 2, 7
	rest	2, –1, 7	0, 0, 0

Figure 4: Jo's examination, payoffs in lifetime utility for Jo_1, Jo_2 and Jo_3; lifetime utility for player *i* is the sum of the experienced utlities of all other players plus two times the experienced utility of player *i*, representing the timeslice's double weighting of its own outcomes

Even if each transient agent cares about the others, a little bit of present bias can still lead to a problem of self-control. Imagine that, as in the O'Donaghue and Rabin (1999) model above, each transient agent values the experiences of all transient agents, but places double the weight on its own experiences as those of other timeslices. Figure 4 shows the payoffs in terms of these preferences over the lifetime. Now some of the benefits of studying accrue to the transient agent who studies (because the payoffs of the other transient agents are in her utility function), but there is still an externality and the costs of studying still outweigh the benefits for each individual transient agent. In this lifetime preference model, Jo_1 and Jo_2 are playing a sequential prisoner's dilemma. The dominant strategy is *rest*, but every transient agent prefers the outcome of *(study, study)* to those of *(rest, rest).* By backwards induction reasoning, Jo_1 can predict that Jo_2 will choose *rest.* So Jo_1's choice is effectively between the sequences (*study*, *rest*) and (*rest*, *rest*); (*rest*, *rest*) is superior from her point of view, so according to decision theory she should choose *rest*.

However, if we allow that each transient agent can ask "What should *I*-the person over time do?", then it may be possible to achieve the outcome (*study*, *study*). Intra-personal team reasoning can solve the intra-personal problem of cooperation in the same

way that inter-personal team reasoning solves the inter-personal problem. Intra-personal team reasoning could proceed according to either of Sugden's or Bacharach's theories.

In the game in terms of lifetime utilities (Figure 4), there is an opportunity for mutual benefit, so we can apply Sugden's mutually assured team reasoning. If Jo_1 has reason to believe that Jo_2 will identify with the team of the person over time and will endorse mutually assured intra-personal team reasoning, then if Jo_1 also endorses mutually assured intra-personal team reasoning, she can choose to *study*. This captures our intuition that starting a plan that will require a series of sacrifices, such as a study plan or a diet, requires the belief that our later self will follow through.

However, we might wonder whether mutually assured team reasoning is the right framework for thinking about the self over time. It is built on ideas of reciprocity and the social contract, which do not seem to apply in the case of the self over time. The lifetime utility is constructed from the timeslices' preferences, with the transient agents compromising on their timeslice preference satisfaction. As well as the basic implausibility of this approach, we might also worry that it introduces an element of double counting into the goals of the person over time. In the realm of social contract theory, Dworkin (1977) distinguished "personal preferences", which are wholly about oneself, and "external preferences", which are about other people. He argues that people's external preferences should not influence the assignment of goods. In the intra-personal case, if Jo_1 is positively disposed towards Jo_2 and wants her to have good outcomes and, as well as allowing Jo_2's personal preferences to influence what the team seeks to achieve we also take into account Jo_1's external preference, then we have double counted Jo_2's outcomes.

We can also apply Bacharach's circumspect team reasoning, since Jo_1 has to make her choice before she knows for sure whether Jo_2 will group identify and team reason. In the Bacharach framework, there is no particular reason to think that the lifetime utility function is based on each transient agent's lifetime preferences rather than on each transient agent's outcomes. As a simplification, let us assume that the lifetime function is achieved by aggregating the transient agents' experienced utility. However, in that case, as this stands, the model would lead to unconditional cooperation (self-control) by any timeslice who team reasons. That is guaranteed by the externality structure of the problem:

the cost born by the timeslice is always outweighed by the benefits to the set of timeslices, so *study* will always be better for the team regardless of whether the other timeslices group identify. In fact, we might even wonder why we need team reasoning. If a timeslice simply takes the aggregated utility of all timeslices as their end, that would suffice to get them to exercise self-control. So this model both violates the intuition that an person will not usually start on a plan unless she expects her later selves to follow through and obviates the need for team reasoning.

We can re-introduce an element of conditional cooperation and, with it, a need for team reasoning if we turn the examination problem into a threshold case. A *threshold public good* does not have a linear relationship between costs and benefits. Rather, the good is provided if and only if contributions pass a minimum threshold. As applied to the problem of Jo's examination, imagine that the exam is pass-fail and Jo needs to study both days in order to pass. The only change we need make to the original problem is in period 3, where experienced utility is 0 unless Jo has worked on both previous days, in which case it is 10. Now the outcome matrix in experienced utilities is as in Figure 5 and the aggregated outcomes when both players view the problem from the perspective of the intra-personal team are in Figure 6. From the perspective of the team, this is a Hi-Lo game. We can see that (*study*, *study)* gives better outcomes for the team than (*rest, rest*), but either of these is better than the outcome where one timeslice works and the other rests. Therefore, if Jo_1 group identifies and team reasons, then what she should do depends on whether or not she expects Jo_2 to group identify. If she expects that Jo_2 will also team reason then she should choose *study*, but if she expects that Jo_2 will not team reason, and will therefore choose the individually dominant strategy of *rest*, Jo_1 should *rest* herself. Whether or not a team reasoning transient agent will exercise self-control will depend on the payoffs involved and on the strength of her belief that later timeslices will also team reason.

		Jo_2	
		study	*rest*
Jo_1	*study*	-3, -3, 10	–3, 0, 0

	rest	0, -3, 0	0, 0, 0

Figure 5: Jo's examination threshold case, experienced utility payoffs shown for Jo_1, Jo_2 and Jo_3

		Jo_2	
		study	*rest*
Jo_1	*study*	4, 4, 4	-3, -3, -3
	rest	-3, -3, -3	0, 0, 0

Figure 6: Jo's examination threshold case, when both players view the problem from the perspective of the intra-personal team and aggregate the transient agents' experienced utilities to obtain the team payoffs

It is not difficult to work out when a team reasoning Jo_1 will choose *study.* We know that if Jo_1 does not study, then the best response by Jo_2 will be *rest* from both the perspective of the timeslice and that of the team over time: choosing *rest* is the unconditional best response from the perspective of the timeslice and it is the best response for a team reasoning Jo_2 if Jo_1 has chosen rest. Remembering that a team reasoning Jo_1 will maximise the payoff of the team and assigning $0 < p < 1$ as the probability that Jo_2 will group identify and U_t as the team payoff function, then Jo_2 will *study* if:

expected team payoff if Jo_1 chooses *study* > expected team payoff if Jo_1 chooses *rest*

$=>$ p(Jo_2 chooses *study*).U_t(*study*, *study*) + p(Jo_2 chooses *rest*). U_t(*study, rest*)> U_t(*rest*, *rest*)

$=> 4p - 3(1-p) > 0$

$=> 7p > 3$

$=> p > 3/7$

Therefore Bacharach-style intra-personal team reasoning, understood as a psychological process of identifying with the person over time, can generate a plausible theory of rational self control (one that is conditional on what later timeslices are expected to do) if the structure of the intra-personal problem is a threshold public goods game. It is not implausible to think that problems of self-control, as viewed by the person involved, have this threshold structure. Although most self-control problems have an underlying continuous public goods game structure, we may have a tendency to turn them into threshold cases. Most self-control problems have continuous but imperceptible benefits. Think of smoking, where every cigarette has a very small negative effect on the smoker's health, or dieting, where every calorie that the dieter forgoes consuming puts her nearer to fitting into that dress. However, when we forgo these temptations, we are looking for perceptible benefits. The smoker wants to feel healthier, the dieter wants to lose a kilo or to fit into a dress. These perceptible benefits fix a threshold—albeit one that is vague—the number of cigarettes or calories forgone to make a perceptible difference. If the person is aiming for a perceptible difference, then there is no point an earlier self forgoing the first cigarette or the first dessert unless she expects enough of the later timeslices to continue the good work.

Of course, the theory of intra-personal team reasoning needs to be supplemented with an account of how timeslices come to identify with the person over time. This is not the place to develop one, but here is a sketch. (I say more about it elsewhere, in Gold, unpublished; Gold,& Kyratsous, 2017). Again we can make an analogy to the inter-personal case. In psychology, there is a body of research about how individuals come to identify with groups. The mechanisms of group identification fit into two broad categories: recognising that the group members have some sort of shared goal or common fate, and recognising commonalities or similarities between the individuals within the group. Both of these could apply to the intra-personal case.

Timeslices may identify with the person over time because they recognise that they all share long-term interests. As Korsgaard (1989) argues, there is a sense in which timeslices are one continuing person *because* they have one life to lead. Her arguments are normative, but a psychological and phenomenological analogue can be found in the work

of James (1890), where one source of a sense of self over time is the recognition that a self in the past or the future was or will be part of the same person. Therefore we might speculate that increasing the salience of the shared interests of the timeslices, or their long-term goals, will facilitate this sense of identification.

Alternatively, timeslices might identify with the person over time because they realise that they are either similar to or connected to the other timeslices. James (1890) also thought that that the current self's perception that it is similar to proximate selves gives rise to a sense that it is continuous with those proximate selves. Psychologists have found that subjects who rated themselves as more connected to later selves, in the sense of Parfit (1984), were more patient (Bartels & Rips, 2010) and that connectedness can be manipulated, resulting in increased patience (Bartels & Urminsky, 2011). Accordingly, increased salience of either similarities or connectedness between the timeslices may facilitate identification with the person over time.

Decision theory provides a model of instrumental rationality, where decision-makers take the best means to their ends. Instrumental rationality presupposes that the decision-maker has a set of ends. Therefore, the timeslice has to identify with a level of agency, and take on a set of ends, before instrumental reasoning can begin. However, decision theory has nothing to say about phenomenology. It is consistent with this picture that the timeslice experiences a tension between between the transient-agent preferences and the self-over-time preferences, so the model is compatible with the phenomenology of conflict.

6. Willpower, decision theory, and intentions

Intra-personal team reasoning sheds fresh light on willpower. In the model, willpower is the ability to align one's present self with one's extended interests by identifying with one's self over time. This picture of willpower differs from the idea of Holton (1999, 2009), that strength of will consists in not reconsidering one's resolutions. But it does create a space for resolutions in decision theory and resolves a puzzle about resolutions that we find in Holton's account.

Standard decision theory does not have room for intentions—understood as motivating plans— or for resolutions that are designed to fortify us against contrary inclinations later on. In O'Donaghue and Rabin's (1999) model, the naive agent who forms a plan in T0 to write the report in T1 will not carry out the plan and ends up missing the Depp movie. Part of her problem was that she did not take into account the preferences that her T1 self would have. A sophisticated agent would correctly predict her future present bias and her T0 self would be able to plan to do the report in T2. However, this is simply a correct prediction of her future behaviour. If prediction is all that planning consists of, then we do not need a separate concept of a "plan" and there is no need for intentions because we can suffice with beliefs. Further, the hyper-rational agents of standard economics can make optimal decisions in a flash; they can make them whenever and how ever many times as they want, so there is no need to make them in advance and form the intention to act on the decision later.

Decision theory can make room for intentions by appealing to the idea of bounded rationality. If it takes time for a boundedly-rational agent to make a decision, then it may not be best for the agent to take the decision at the time of action. Once the person has made a decision then, other things being equal, she should not waste her limited time by re-opening the question. Hence, it might be optimal for the agent to make the decision in advance and form an intention as a reminder of her decision, which she can consult at the time of action.

However, the idea of a resolution fits uncomfortably in this framework. Remember that agency sits with timeslices. Effectively, the past timeslice takes a decision and, other things being equal, the future timeslice does not re-open the question. (If decision making is onerous, then there may also be a problem of procrastination about making the decision, but we leave that aside here.) One part of "other things being equal" is the idea that the later timeslice would be likely to make the same decision as previously, so re-deliberating is a waste of time. This seems uncontroversial in cases where there is no conflict of interests between timeslices. However, Holton's (1999, 2009) resolutions are formed in order to defeat contrary inclinations, which will arise at the time of action. In the decision theoretic framework, the past timeslice is making a decision that she knows the future

timeslice will not want to carry out, if the future timeslice takes the standard timeslice perspective. Therefore, if a timeslice remembers that an earlier timeslice made a resolution, she also has reason to think that the resolution conflicts with her current timeslice-preferences, so if she is thinking as a timeslice then she should abandon any prior resolutions. (This relates to Hinchman's 1993 idea that diachronic agency involves a type of self-trust.)

If the timeslice is to act on resolutions, then we need to add something extra and intra-personal team reasoning can supply the missing piece of the puzzle. In the framework of the person as a team over time, an intention is a plan made by an earlier timeslice who identifies as a member of the team over time. The intention has two different purposes. Firstly, it is a contingency plan, for later timeslices who turn out to identify with the person over time and therefore share the team preferences of the planner. There is no conflict of interest and, if the later timeslice has no reason to suspect that the earlier planning timeslice made a bad plan, then she can simply follow her part of the plan. Intra-personal team reasoning can explain how the different timeslices' interests are aligned so that the later timeslice knows that she should follow the plan made by the earlier one.

Planning can also play a second type of role in this picture. In the same way that standard decision theory allows earlier timeslices to take actions to constrain later timeslices, in the theory of intra-personal team reasoning the earlier timeslice can take actions that increase the probability of group identification by later ones. Remembering a plan may encourage the timeslice that does the remembering to identify with the person over time. For instance, it makes salient the existence of the temporally extended agent and the shared extended interests of the timeslices.

In this second role, making a plan may have some of the effects of Holton's (1999, 2009) resolutions. By encouraging the later timeslice to identify with the person over time and therefore to act on the plan, it may prevent the transient-agent reasoning that leads to weakness of will. Therefore resolutions are a mechanism of self-control. Nevertheless, in the model of intra-personal team reasoning, the resolution is not the root cause of self-control. The fact that the plan can be effective in the face of contrary inclinations is parasitic on the idea that the timeslice can identify with the self over time and do intra-

personal team reasoning. If, at the time of action, the timeslice did some reasoning then she could come to the same decision as previously, provided that she does intra-personal team reasoning rather than transient-agent reasoning. Further, an agent who makes a resolution but then happens to reconsider at time of action is not totally lost (so intra-personal team reasoning solves a problem posed by Bratman, 2014, about how an agent can rationally form an intention if she anticipates that she will re-open the question and take a transient-agent view at the time of action). It is not a foregone conclusion that the later timeslice will do transient-agent reasoning rather than intra-personal team reasoning.

We can also compare willpower as resolutions to willpower as intra-personal team reasoning using the philosophical framework of synchronic versus diachronic self-control (Mele, 1987). *Diachronic self-control* occurs when an agent anticipates a preference change and takes action to prevent herself from succumbing later, *synchronic self-control* occurs when the agent exercises self-control at the very same time as experiencing a temptation. Intentions can be a means of diachronic self-control in both the resolution and the intra-personal team reasoning accounts. In the resolutions account, this is because not re-considering one's resolution is the instrument of synchronic self-control. In the intra-personal team reasoning account, an agent who identifies with the person over time might not re-consider her intentions. However, the ultimate instrument of synchronic self-control, which also underpins the intention or resolution, is intra-personal team reasoning. If an agent forms a resolution, it is effective because it prompts identification with the person over time and, hence, acting on the results of intra-personal team reasoning.

7. Conclusion

I have presented a picture of willpower as intra-personal team reasoning, analogous to using inter-personal team reasoning to solve problems of cooperation between individuals. I suggested that we should model problems of self-control as threshold public goods games. I have shown how, although the timeslices' transient-agent preferences are in conflict, it is possible for them to identify with the person over time and use circumspect intra-personal team reasoning in order to resolve their problem of self-control. Intra-personal team reasoning also provides a basis for introducing intentions and resolutions

into decision theory, although at base it is intra-personal team reasoning that solves the synchronic problem of self-control and which gives resolutions their power.

I have shown how willpower can be instrumentally rational for the person over time, even while succumbing to temptation is instrumentally rational from the perspective of the timeslice. In this sense, the model provides an answer to the longstanding philosophical question of how an individual can intentionally act against what she judges to be best. Many people have the intuition that the timeslice is doing something wrong if she succumbs to temptation. It follows from what I have said here that the wrongness is not derived from instrumental rationality. There is a lot more to be said about why timeslices should identify with the self over time and when they will do so (Gold, unpublished). But this is an issue for a whole separate paper.

References

Ainslie, George (1992) *Picoeconomics* (Cambridge University Press: Cambridge)

Bacharach, M. (1997): ''We' Equilibria: A Variable Frame Theory of Cooperation', Oxford: Institute of Economics and Statistics, University of Oxford, 30.

Bacharach, M. (1999): 'Interactive Team Reasoning: A Contribution to the Theory of Co-operation', *Research in Economics*, 53, pp. 117-147. inmore, Ken (1987) 'Modelling rational players: Part I'. *Economics and Philosophy* 3, 9-55.

Bacharach, M. (2006): Beyond Individual Choice: Teams and Frames in Game Theory, Princeton: Princeton University Press.

Bartels, D. M., & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General*, *139*(1), 49-69.

Bartels, D. M., & Urminsky, O. (2011). On intertemporal selfishness: How the perceived instability of identity underlies impatient consumption. *Journal of Consumer Research*, *38*(1), 182-198.

Bratman, M. E. (2014). Temptation and the Agent's Standpoint. *Inquiry*, *57*(3), 293-310.

Colman, A. M., & Gold, N. (2017). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review*, 1-14.

Van Boven, L., Loewenstein, G., & Dunning, D. (2005). The illusion of courage in social prediction: Underestimating the impact of fear of embarrassment on other people. *Organizational Behavior and Human Decision Processes, 96(2)*, 130-141.

Dworkin, R. (1977). Taking rights seriously (Vol. 136). Cambridge Ma: Harvard University Press.

Gold, N. (2013). Team Reasoning, Framing, and Self-control: An Aristotelian Account. In N. Levy (ed.) *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*. Oxford: Oxford University Press.

Gold, N. (2012): 'Team Reasoning, Framing and Cooperation', in S. Okasha and K. Binmore (eds), 'Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour', Cambridge: Cambridge University Press, ch. 9, pp. 185-212.

Gold, N. (unpublished manuscript) Guard Against Temptation.

Gold, N. & Kyratsous, M. (2017). Self and identity in Borderline Personality Disorder: Agency and mental time travel. *Journal of Evaluation in Clinical Practice 23(5)*, 1020–1028.

Gold and Sugden (2006) Conclusion. *Beyond Individual Choice.* By Michael Bacharach. Princeton: Princeton University Press.

Gold, N. and Sugden, R. (2007a): 'Collective Intentions and Team Agency', *Journal of Philosophy*, 104, pp. 109-137.

Gold, N. and Sugden, R. (2007b): 'Theories of Team Agency', in F. Peter and H. B. Schmid (eds), 'Rationality and Commitment', Oxford: Oxford University Press, pp. 280-312.

Greer, J. M., & Levine, D. K. (2006). A dual-self model of impulse control. The American Economic Review, 96(5), 1449-1476.

Hinchman, E. S. (2003). Trust and diachronic agency. *Noûs*, *37*(1), 25-51.

Holton, R. (1999). Intention and weakness of will. The Journal of Philosophy, 96(5), 241-262.

Holton, R. (2009). Willing, wanting, waiting. Oxford University Press.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. American Economic Review, 103-124.

James, W. (1890). *The Principles of Psychology.* New York: H. Holt and Company.

Kahneman, D., & Thaler, R. H. (2006). Anomalies: Utility maximization and experienced utility. *The Journal of Economic Perspectives*, *20*(1), 221-234.

Karpus, J & Gold, N. (2016). Team reasoning: Theory and evidence. in J. Kiverstein (ed) *Handbook of Philosophy of the Social Mind*. pp.400-17. New York: Routledge.

Korsgaard, C. M. (1989). Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy & Public Affairs*, 101-132.

Loewenstein, G., O'Donoghue, T. & Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics, 118*, 1209-1248.

Lowenthal, D. & Loewenstein, G. (2001) Can voters predict changes in their own attitudes? *Political Psychology, 22*(1), 65-87.

Mele, A. R. (1992). *Irrationality: An essay on akrasia, self-deception, and self-control*. Oxford University Press.

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

Pettit, Philip and Robert Sugden (1989) 'The backward induction paradox' *Journal of Philosophy* 86: 169182.

Pronin, E., & Ross, L. (2006). Temporal differences in trait self-ascription: when the self is seen as an other. *Journal of personality and social psychology*, *90*(2), 197=209.

Reny, Philip (1992) 'Backward induction, normal form perfection and explicable equilibria' *Econometrica* 60, 627-649.

Schelling, T. C. (1984). Self-command in practice, in policy, and in a theory of rational choice. The American Economic Review, 74(2), 1-11.

Sugden, R. (1993): 'Thinking as a Team: Towards an Explanation of Nonselfish Behavior', *Social Philosophy and Policy*, 10, pp. 69-89.

Sugden, R. (2000): 'Team Preferences', *Economics and Philosophy*, 16, pp. 175-204.

Sugden, R. (2003): 'The Logic of Team Reasoning', *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 6, pp. 165-181.

Sugden, R. (2011): 'Mutual Advantage, Conventions and Team Reasoning', *International Review of Economics*, 58, pp. 9-20.

Sugden, R. (2015): 'Team Reasoning and Intentional Cooperation for Mutual Benefit', *Journal of Social Ontology*, 1, pp. 143-166.

Strotz, Robert H. (1955-6) `Myopia and Inconsistency in Dynamic Utility Maximization', *Review of Economic Studies* 23, 165-80.

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. Journal of political Economy, 89(2), 392-406.

Acknowledgements

Work on this chapter was supported by funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 283849. The author thanks the project team, Jurgis Karpus and James Thom, for helpful discussions. She also thanks the participants at the Texas A&M workshop on Rationality and Self-Control for their feedback on this paper, especially José Luis Bermúdez who provided written comments, and the participants at the Stanford workshop on Varieties of Agency, especially Michael Bratman and Grace Paterson, whose feedback on a companion paper also had the side-effect of improving this one.