DOI: 10.1111/nous.12369

ARTICLE

Noûs

Guard against temptation: Intrapersonal team reasoning and the role of intentions in exercising willpower

Natalie Gold

Centre for Philosophy of Natural and Social Science, London School of Economics

Correspondence
Natalie Gold, Centre for Philosophy of Natural and Social Science, London School of Economics.
Email: n.gold@lse.ac.uk

Forthcoming in *Nous*

I am grateful for in-depth discussion with the 'Group Thinking: New Foundations' European Research Council (ERC) project at Leeds (grant agreement no. 818633) and at the Stanford 'Varieties of Agency' series, where Grace Paterson was an insightful commentator. A companion paper, Gold (2018) 'The person as a team over time', focusses on the decision theoretic structure of the problem. My work on this paper 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.

Abstract
Sometimes we make a decision about an action we will undertake later and form an intention, but our judgment of what it is best to do undergoes a temporary shift when the time for action comes round. What makes it rational not to give in to temptation? Many contemporary solutions privilege diachronic rationality; in some "rational non-reconsideration" (RNR) accounts once the agent forms an intention, it is rational not to reconsider. This leads to other puzzles: how can someone be motivated to follow a plan that is contrary to their current judgment? How can it be rational to form a plan to resist if we can predict that our judgment will shift? I show how these puzzles can be solved in a framework where there are multiple units of agency, distinguishing between the judgments of the timeslice and those of the person over time, and allowing that the timeslice can "self identify", taking the person over time as the relevant unit of agency and doing intrapersonal team reasoning (with a different causal role for intentions than RNR accounts). On my account, resisting temptation is compatible with synchronic rationality, so synchronic and diachronic rationality are aligned. However, either

resisting or succumbing to temptation can be instrumentally rational, depending on the unit of agency that is identified with. In order to show why we ought to resist temptation, we need to draw on a non-instrumental rationale. I sketch possible routes for doing this.

KEYWORDS
akrasia, decision theory, diachronic rationality, intentions, team reasoning, weakness of will

1 | PHILOSOPHICAL PUZZLES: INTENTIONS IN THE FACE OF TEMPTATION

Sometimes we make a decision about an action we will undertake at a later date and form an intention to do it, but our judgment of what it is best to do undergoes a temporary shift when the time for action comes round. We may have decided in advance to forgo something nice, for instance I may decide that I will decline the offer of dessert after dinner because I am on a diet. Or we may have decided to do something that will be unpleasant, for instance I may decide to go on a run tomorrow morning, which I know I will enjoy, but that will require getting up early, which will be painful at the time. Actions like refusing dessert or getting up at the crack of dawn may not seem like such a good idea when the time comes to perform them. After the time of action, our judgements revert. If we have succumbed to temptation then we will regret it.

There are also more complex problems, where the temptation is embedded in a more complicated plan. For instance, under forgoing pleasures we have what I call *stopping problems*. We want a small amount of some pleasure, but we know we will struggle to stop once we have started: I would like to eat just one biscuit with my cup of coffee, to read just one chapter of my pacey novel, or to go online for just a few minutes. If we binge, then afterwards we will regret it. Under avoiding displeasures, we have *scheduling problems*. We have to do some unpleasant task and we work out the optimal time to do it, but when the time comes we are tempted to do something else instead. For instance, I may have decided to do my marking on Saturday morning, so that I don't have to worry about it all weekend, but on Saturday I judge that my morning could be better spent doing some DIY, and I end up doing my marking on Sunday night, which means I miss my weekly movie night and I regret my procrastination. Although procrastination may seem the obvious outcome, a scheduling problem can in theory also lead to someone doing a task earlier than would be optimal (O'Donaghue and Rabin, 1999).

The common structure of these problems is that there is a choice of action or activity, and the option that we judge to be best before and after the time of action is different to the one that is judged best at the time of action. We experience a reevaluation of the options that Richard Holton (2009) refers to as a *judgment shift*. (In the decision theoretic literature, this is called a *preference reversal*, but "preference" is a catch-all term to represent why one option is ranked higher than another and there is no implication that the rank ordering is based on what a philosopher might think of as mere preferences.) This is more than just feeling the pull of temptation. As Bratman (2014) puts it, when he experiences temptation he will change how he values the options "in light of considerations that now matter to me" (Bratman 2014, p.298). The judgement shift is temporary and can be predicted in advance, so these cases do not involve any new information over the course of time.

Let us ground our discussion in an example. Following Bratman (1999, 2014), I will use a stopping problem. However, Bratman's example involves wine, which could be thought to cloud rational judgment, so I will stick to biscuits. I would like to eat just one biscuit with my cup of coffee at teatime, but I know that, if I have one then I'll want another… and another. Effectively the same "just one more" decision structure occurs every time I eat a biscuit, until I get to the end of the pack. After I've scoffed the whole pack, then I'll regret it and wish I'd stopped at one—and I even think that it would have been better not to have started eating biscuits at all than to have had more than one. One thing I might do in advance of teatime is to form an intention to only have one biscuit with my coffee. However, at the time I form the intention, I also know that at teatime I will be tempted to have a second biscuit. This leads to a series of puzzles.

Puzzle 1. Come teatime, what makes it rational not to eat a second biscuit? One answer grounds the rationality of abstaining in the special character of intentions. A prominent proponent of this position is Holton (1999), who argues that there is a norm of rationality that one should not revise one's intention when standing firm "manifests tendencies that it is reasonable for the agent to have" (p.6). John Broome (2013, p.178) also offers an account whereby it is a requirement of rationality that intentions persist until or unless a "cancelling event" occurs. Let us follow Bratman (2014) in calling these *Rational Non-Reconsideration*, or RNR, accounts. For Holton, RNR is an important component of willpower. On his account, one reason for forming an intention is precisely to overcome contrary inclinations that we might have at the time of action. These *resolutions* block us from re-opening the question, hence guarding against acting on the judgment shift.

However, there are several issues with RNR accounts.

While RNR accounts can tell us why it is rational not to re-open the question, they offer no tools once the question has been re-opened. Psychologically it may be hard to avoid re-considering intentions in the face of temptation (Bratman, 2014; Paul, 2011). Holton incorporates this phenomenology into his account, with willpower involving the mental effort of not reconsidering. However, once the question is re-opened, RNR accounts do not have any resources left for the agent to exert self-control, neither an account of rational self-control nor any psychological resources for the agent. For Broome (2013), re-opening the question is a cancelling event. Further, once the question is re-opened, the prior intention has no rational sway in the re-consideration. If the prior intention provided a reason for acting, there could be illicit bootstrapping: forming the intention to do something would give one a reason to do it, even in cases where no reason existed prior to forming the intention. Holton also agrees that, although it is rational not to reconsider a resolution, once the question is re-opened then it would be rational to change one's mind (Holton, 2004). RNR accounts retain what Bratman calls *rational priority of present evaluation*, "if the agent does indeed have a relevant judgment at t2 concerning which alternative would be strictly best at t2, then if that agent is functioning rationally she will opt for that alternative" (Bratman, 2014, p.297). Therefore, if I do reconsider at teatime, then it will be rational to have the second biscuit.

This leads to *Puzzle 2.* There is something puzzling about the agent's motivation not to reconsider in RNR accounts. For Holton (2009), the specific purpose of resolutions is to overcome contrary inclinations at the time of action, so let's think about what this might look like. At teatime, my companion says "Would you like a second biscuit?" and I feel an inclination to have a second biscuit but, because I have a resolution, instead of weighing up whether or not to have a biscuit, I simply answer "No." Phenomenologically, this seems like a plausible account. However, rationally, as soon as I realise that I have a resolution not to have a second biscuit, I can also infer that my judgment has probably shifted in favour of having a second biscuit. There is something strange about the position that it is rational not to revise the resolution, while being in possession of facts that make it easy to infer—indeed, I think it would be quite difficult for me to stop myself from inferring—that my judgment has probably changed.

Puzzle 3. Finally, we have the puzzle that Bratman (2014) used his example to motivate: how can I rationally form the prior intention not to have a second biscuit if I know that, when the time comes to do as I intend, it will be rational for me not to follow through on that intention? Classical decision theory, which we can think of as formalizing instrumental rationality, epitomises the problem. It models agents who choose according to their current evaluations and regards an agent who forms intentions in the hope of getting her future self to act against its present evaluation as naive: forming an intention is at best ineffective and can even be counter-productive (O'Donaghue and Rabin, 1999). When he discussed the puzzle, Bratman argued that this case requires us to "abandon the standard [expected utility] model" (1999, p.42). However, I will argue that we can resolve the puzzles by extending, rather than abandoning, decision theory.

The root of these puzzles is that, in order to guard against temptation, RNR accounts draw on requirements of *diachronic rationality*, which connect attitudes at one time with attitudes at another. However, these come into conflict with requirements of *synchronic rationality*, which apply to the attitudes a person has at a single time. I will approach the puzzles from the other direction, starting with requirements of synchronic rationality that may help the agent to guard against temptation, and building up a solution from there, in which it is possible for synchronic and diachronic rationality to align.

2 | DECISION THEORETIC PUZZLES: INTENTIONS AND THE SELF OVER TIME

In philosophy of action, it is generally accepted that agents can form intentions and make resolutions in order to guard against temptations. In contrast, decision theory does not recognise intentions as mental states that have motivating power. I will show how filling this lacuna in decision theory can also help with our philosophical puzzles about the rationality of acting on and revising intentions.

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. Indeed, although there might be an infinite number of possible timeslices, models confine themselves to the handful that make pivotal decisions.

If different timeslices have different preferences, then one can think of successive timeslices as involved in strategic interaction (Schelling, 1984). The term "preferences", as used in decision theory, is a catch-all term, implying an ordering over the options but nothing about the reasons behind the ordering. So it should not be thought of as modelling "mere preferences" and it incorporates the action theorists' "judgments". In the stopping problem, my preferences at the start of teatime (*time1*) and after it is over (*time3*) will be one biscuit $\succ$ no biscuits $\succ$ two biscuits. However, if I act on these preferences at *time1*, then after eating the first biscuit (*time2*), my preference will be to have a second, giving an ordering of two biscuits ≻ one biscuit ≻ no biscuits. Call these the preferences of three timeslices T1, T2, and T3. The timeslices' preferences are not aligned.

Forming an intention to stop at one biscuit would be naive, as that fails to recognise that T2's preferences will be different from T1. It would be ineffective and result in T2 eating two biscuits, which is worse than other outcomes that T1 could have ensured. According to decision theory, eating just one biscuit isn't an available outcome. If T2 is allowed free rein to act on her preferences, then she will eat a second biscuit. Therefore, reasoning by backwards induction, T1 can see that the highest attainable option in her preference ordering is no biscuits, so she should not eat the first one. This outcome is perverse, since all timeslices prefer one biscuit to none.

The one biscuit outcome is attainable if we add in a T0 timeslice, who has the ability to change either the incentives or the set of options that are faced by T2. For example, she could make a bet with a colleague that she will only eat one biscuit at teatime, so that eating a second biscuit would lead to a loss of money. She may not even need to find someone who will take the other side of a bet, if she cares about loss of face then a public announcement in the common room may suffice. Alternatively, T0 can manipulate the options that are available at T2, for example giving away the rest of the biscuits in the packet, so that there is only one biscuit left for her. (Limiting the current availability of biscuits is arguably another way of increasing the costs of eating a second biscuit, since T2 could presumably still choose to go to the shops and purchase some more. A similar point could be made about limiting the availability of any option that can be replenished.) All of these are possible plans that T0 can implement, in the knowledge that they will be effective: they are *incentive compatible*, as each timeslice's part in the plan will be the preferred action at the time of choice.

However, what T0 cannot do in this framework is simply form an intention to stop at one biscuit, which will itself affect T2's action and guard against temptation. Standard decision theory does not allow plans that have causal power, it only allows the concept of a plan as an incentive-compatible set of future decisions. (This is sometimes known as "sophisticated choice".) T0 could have a plan to give away the excess biscuits, so that T2 won't be able to eat more than one, or T0 could simply have a plan for T2 to eat lots of biscuits. What T0 cannot have is a plan that will motivate T2 not to eat a second biscuit, simply in virtue of having been formed—in other words an intention.

3 | THE PERSON AS A TEAM OVER TIME: SOLUTION TO PUZZLE 1

Intertemporal strategic interactions are similar to interpersonal strategic interactions: at their heart they are about the distribution of costs and benefits between (transient) agents. In economics, there is an *externality* when a person's action has effects on other people that are not captured in his or her own payoffs. We can think of giving into temptation as imposing an externality on future selves (Gold, 2013), sometimes called an *internality* (Read, 2001). It benefits the current self at the expense of imposing costs on later selves. If T2 has a second biscuit, then she gets a benefit but T3 will bear a cost when she feels slightly sick and other selves further down the line may bear any health costs from eating lots of biscuits. Conversely, if T2 stops at one biscuit, then she will bear a cost (forgoing the pleasure of the second biscuit) for benefits that accrue to other timeslices (T3 will not feel sick, later timeslices will be healthier).

The decision theoretic outcome of succumbing to temptation does not require that T2 be completely selfish and have no care for the effects of her action on other timeslices. Indeed, some of the standard economic incentive-compatible solutions to temptation, such as side bets, rely on T2 caring about what happens to T3. A plausible model gives the timeslice some degree of *present bias*, where the timeslice gives its outcomes more weight than those of other timeslices. Interestingly, altruism for other timeslices alone does not necessarily lead to the resistance of temptation. Even a golden-mean altruist timeslice, who is equally concerned with all timeslices, does not necessarily rationally resist temptation when that makes the timeslices better off in aggregate—that depends on the structure of the payoffs.

Temptation is similar to a public goods game, played sequentially over time. A *public goods game*, or *social dilemma*, involves a conflict between individual incentives to free ride and social incentives to contribute toward the provision of a public good: individual contributions have an externality on other players. This has parallels with the prisoner's dilemma, where standard decision theory tells players to defect, even though they would all prefer the outcome where everyone cooperates. In the stopping problem, there is a perverse outcome that the rational-backwards induction solution (no biscuits) is *dominated,* since all timeslices preferred another option (one biscuit). The multi-player version of the prisoner's dilemma is a well-known example of a public goods game, but there are many different *production functions* that can govern the relationship between individual contributions and social outcomes, including *threshold models*, where the public good is only provided if more than a certain number of individuals contribute (Ledyard, 1995; Schelling, 1978).

Ainslie (1992) and Gold and Sugden (2006) have argued that temptation can be seen as a prisoner's dilemma over time. If resisting temptation imposes a cost and the corresponding benefit is spread amongst all timeslices, then every timeslice may have an incentive to succumb, although every timeslice would prefer the outcome of *all resisting* to *all succumbing*. However, in models of self-control, the benefits may not accrue to all timeslices. For instance, in Bratman's puzzle, it is arguable that if T2 stops at one biscuit, then T2 is made worse off by its action, being sacrificed for the benefit of later selves. Alternatively, the benefits may not follow a continuous distribution. Gold (2018) has argued that temptation is better viewed as a threshold public goods game, where at least a certain number of timeslices need to contribute in order to reach a goal that makes the aggregate better off, such as passing an exam, fitting in a dress, or just feeling fitter, healthier and more productive. Those papers also related team reasoning to self control, but they focus on the decision theoretic structure of the problem; here I am interested in interpretation of the model and the normative rationale. For my purposes, we don't need to adjudicate the structure of the game, we only need to agree that we can be in a position where we can say that a timeslice has incentives to behave in one manner but the aggregate of time slices would be better off if that timeslice behaves in a different manner. This is analogous to the way that, in an interpersonal prisoner's dilemma, although each individual individual is better off defecting, they would collectively be better off if they cooperated. Although the answer to the question "what should *I* do?" is "defect", the answer to the question '"what should *we* do?" is clearly not "all defect", since the players would be better off if all cooperated, but standard game theory does not admit the collective question.

The theory of team reasoning extends the syntax of game theory to allow players to ask "what should *we* do?", which admits "all cooperate" as an answer (Gold & Sugden, 2007). 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 (What should *we* do? All cooperate) and then performs her part of that combination (How can *I* play my part? Cooperate). We can apply the theory of team reasoning to intertemporal choice, thinking of the person as a team-over-time. In decision theoretic models of intertemporal choice, each timeslice asks separately "What should *I-now* do?". The emphasis on the timeslice gives a natural model of succumbing to temptations. But it is also possible for the timeslice to ask, "What should

I-over-time do?" and "How should *I-now* play my part?", doing what would best promote the interests and objectives of the long-term self. For many timeslices, changing the question will not give a different answer. However, in cases of temptation, given the public goods game structure, the answers may come apart. Asking the question from the perspective of the person over time opens up the possibility that the timeslice should resist. For instance, in my example, it is arguable that it is best for the person over time if T2 only has one biscuit, so if T2 does intra-personal team reasoning, she can conclude that she should not have a second biscuit.

The idea of the person as a team over time involves more than just caring about the outcomes of past and future timeslices, or *intrapersonal altruism* (Gold, 2013). Team reasoning involves an *agency transformation*, in which the agent identifies with the team agent and its goals, and acts as a sub-part of the agent. This is not just about the phenomenological properties of the experience, of caring versus identification. Intuitively, it may seem that, if caring takes the form of an equal concern for all timeslices, this should be sufficient for cooperation. However, cooperation also depends on the structure of the game: if the relationship between costs and benefits is not linear, then intrapersonal altruism may not be sufficient. For example, interpersonal altruism is not sufficient for resistance if the overall structure is that of a *threshold public good*, with a desirable outcome that is only achieved if some minimum number of timeslices resist temptation, e.g., reaching a pass mark on an exam that requires at least some minimum number of hours of study or losing weight to fit into a dress that requires dessert to be skipped some minimum number of times (Gold, 2018). We might think of this as introducing an element of intrapersonal coordination. In these cases intrapersonal altruism does not recommend unconditional resistance; for that we need identification and the agency transformation involved in intrapersonal team reasoning.

This solves Puzzle 1: Team reasoning can explain why it is rational for me to resist temptation, if I re-open the question at teatime. If I identify with the person over time (from hereon, the person), then it is rational to team reason and to play my part in the plan that is best for the person. Note that this is an entirely synchronic solution, there is no reference to intentions or to diachronic norms of rationality. However, it might be objected that I have merely pushed the question back one stage: why identify with the person? I will return to that question later. First, we need to think about what it is to self identify and how to understand the relationship between timeslices, and between timeslices and the person.

4 | THE RELATIONSHIP BETWEEN TIMESLICES AND SELF IDENTIFICATION WITH THE PERSON

Before pursuing the intrapersonal team reasoning account of self-control further, we need to address a concern about whether there is a sensible account of timeslices that can underpin this picture. By providing an account, I will also solve Puzzle 2.

Thoma (2018) argues that timeslice accounts are caught between two ends of an unappealing spectrum. At one end of the spectrum, we recognise that they are timeslices of the same agent, in which case why wouldn't the later preferences completely override the earlier ones? At the other end, the timeslices are completely separate, in which case any cooperative scheme is forced on the current timeslice, who never agreed to it. In order to avoid this Scylla and Charybdis, I start by noting that the intra-personal team reasoning account of the person includes two levels of unit of agency, the timeslice and the person. (In this context, it makes sense to talk of levels because the timeslices are nested matryoshka-like, within the person, which in turn is nested in larger groups or teams of people.) Then the dilemma can be averted by differentiating between an account of how timeslices relate to each other and an account of how they relate to the person.

Timeslices may be separate from each other, but this does not imply that their concerns are completely separate from each other, and there are a couple of ways that we could conceptualise their relationship to each other.

Parfit (1984) introduced the idea of *psychological connectedness* between timeslices, such as shared memories, beliefs, and desires, as a part of his theory of personal identify. For Parfit, the fact that a timeslice is only weakly psychologically connected to future timeslices is a reason for rationally caring less about the interests of one's future self. The less connected one is, then the less one need care. Psychologists have investigated the descriptive version of this claim: that as connectedness to future selves decreases, the less people *in fact* care about about their future selves. They find that people who rate themselves as more connected to their future selves are more patient and willing to wait for a distant reward (they discount future benefits by a smaller amount), and that this sense of connectedness can be manipulated and, when it is manipulated to be lower, people are more impatient and have a higher discount rate (Bartels & Rips, 2010; Bartels & Urminsky). This looks like the standard economists' models of intertemporal choice with time-discounting (and no preference reversals), where the current timeslice uses a constant discount rate on the outcomes of future selves.

Another way of conceptualising the relationship between timeslices is to think of the current timeslice as giving itself greater weight, but giving equal weight to all the other timeslices (Gold and Sugden, 2006; Pettigrew, 2019). Pettigrew (2019) fills in the details by making an analogy to the relationship between an individual and a collective, such as an activist group: "In these collectives, executive power passes from one member to another for a specified time period: Sarah is in charge for one month, then Adil, then Cleo, and so on. The individual in charge can place some extra weight on their own preferences, but they must also give weight to the preferences of the others in the collective." In these examples, the timeslice puts equal weight on all the other timeslices, only giving greater weight to itself. This can lead to a model of hyperbolic discounting, where the discount rate changes depending on how close the reward is, which can lead to a change of preferences: someone who prefers a larger-later reward to a smaller-sooner reward reverses and chooses the smaller-sooner reward as the time for receiving the small reward approaches (Gold & Sugden 2006).

These are two different ways that a timeslice's preferences can be concerned with other timeslices' outcomes.

With either option, we can extend Pettigrew's (2019) analogy of collective self-government in order to explain the relationship between the current timeslice and the person. We can think of the timeslices as analogous to a nation with a parliamentary system, except that instead of being elected the powers of the executive pass from one timeslice to another in turn. The executive (the current timeslice) is supposed to use judgment to further the interests of all its constituents (the timeslices that make up the person), deliberating and acting in the common good. When a timeslice makes a law (an intention), then the law remains in force unless it is repealed. However, the executive is sovereign and can override previous decisions, repealing laws (overturning intentions). The executive also has the power to sign treaties and enter into legally binding agreements with third parties (make commitments, such as promises, and incur obligations to others), on behalf of future governments and constituents. On this picture, identifying with the person and doing intrapersonal team reasoning is akin to the "remarkable change in man" that Hollis (1998) attributes to individuals in his Rousseauian account of how people behave in groups. In the same manner that the nation persists even though individual citizens die, the person persists even though individual timeslices are fleeting. Both the nation and the individual can see their interests in the context of their past history and have vexing choices about when to revisit or override past preferences, of past citizens or timeslices who are no longer in existence; both can bind future members who are not yet in existence.

For interpersonal team reasoning to occur, an individual must group identify. An individual is said to identify with a group if she conceives of the group as a unit of agency, acting as a single entity in pursuit of some single objective (Gold & Sugden, 2007). This elides two potentially separate steps, classifying oneself as a part of a group and taking on its commitments, and leaves the mechanisms by which the steps occur entirely open (Gold, 2012). Different accounts of team reasoning fill them in in different ways, ranging from an automatic psychological mechanism to a conscious choice. I will be neutral between mechanisms here, though some mechanisms that have been proposed for the interpersonal case seem less plausible for the intrapersonal (Gold, 2018). Since the mechanism of identification in the intrapersonal case may be different from the interpersonal, let us refer to the timeslice as *self identifying* with the person.

The two ways that the timeslices can experience the relationship between timeslices correspond to two ways that they may come to self identify, which correspond to two different psychological mechanisms postulated by William James. James (1890) thought that the current self's perception that it is similar to proximate selves gives rise to a sense that the current self is continuous with those proximate selves. Having a sense that one is continuous with other timeslices may lead one to identify with them. But James also thought that there is second form of self identification, which consists of the recognition by the current self that a self in the past or the future was or will be part of the same person. He clarified this with a metaphor: The owner of a herd of cattle can recognise his animals because they are branded. However, "no beast would be so branded unless he belonged to the owner of the herd. They are not his because they are branded; they are branded because they are his." (James, 1890 p.337). The selves at different times, which are equivalent to the cattle in the metaphor, can recognise that they are branded the same; being branded the same derives from being a part of the same entity over time. James thought that the semantic understanding that an action or experience was or will be mine is typically accompanied by a feeling of "mineness", as the self projects itself into the past or the future.

Now to return to the questions with which we started this section. At the timeslice-level of agency, we can consider timeslices as separate agents who may reason as timeslices, though this does not imply that they are unconcerned about each other. We can grant that, qua timeslices, their current timeslice preferences replace and over-ride earlier timeslice preferences. However, there is a sense in which the timeslices are a part of the same higher-level agent, the person. If they recognise that they are all timeslices of the same person and they self identify, then they can use intrapersonal team reasoning to pursue the ends of the person.

This solves Puzzle 2, about the agent's motivation not to reconsider in RNR accounts, even though they can infer that a judgment shift will have occurred. The intrapersonal team reasoning account holds that it is rational not to revise the resolution *if the timeslice identifies with the person.* When that happens, then the timeslice will see her interests as those of the person and that can motivate compliance. She may also realise that her judgment and interests qua timeslice would differ from her judgments and interests qua the person, but she puts her interests qua person first, in the same way that, in the interpersonal case, individuals can identify with groups and act in the group interest, even when the group interests come apart from their own. Of course there is a question that can be asked in both the inter- and the intra-individual case, about the extent to which lower- and higher-level interests can come apart without imperilling the identification with the higher level of agency. But there is not space to pursue that here.

In my explanation of temptation, there is no change of mind at the person-level: the optimal team plan (or best judgment or person-level preference, depending on your disciplinary vocabulary) does not change over the relevant time period. The temporary judgment shift occurs at the timeslice-level of agency, and the problem of self-control is caused because the timeslice and the person-level judgments come apart. The change of timeslice preferences may even be predictable and have been taken account of in the person-level evaluation. The problem of temptation is resolved if the timeslice identifies with the person and therefore with the person-level judgment.

There is an entirely separate—but big and important—question of how to think about changes of mind at the level of the person. This has been given a book-length treatment by Pettigrew (2020). However, I will quickly note that the idea of different units of agency at the different levels introduces an additional wrinkle. If someone can be making judgments from the perspective of the timeslice or from the perspective of the person, but there is no neutral "view from nowhere" from which to survey the options, then there is the potential for self-deception, where a timeslice whose judgments are out of sync with that of the person persuades themselves that it is their enduring person-level judgments that have changed.

5 | TEAM MECHANISMS

In the paradigm examples of interpersonal team reasoning, each individual uses team reasoning to decide what to do. In the intrapersonal case, this seems too demanding: not only is it too demanding to expect every timeslice to *team* reason, it is too demanding to expect each timeslice to *reason*. Luckily, both these problems of overdemandingness can be solved by modifications of the theory that already exist in the interpersonal case. For this, we need the idea of a team mechanism, which will also help solve Puzzle 3.

Bacharach (2006) defines a *choice mechanism* as a causal process that determines what people do. A *team mechanism* is one is that ensures the team's common goal is achieved (at least in the usual scheme of things, since the reasoning is always ex ante). Working out what team members should do consists of three steps: computing the best profile of actions for the team, identifying the component of the profile that falls to any particular individual, and telling that individual what her component is. Then the individual agents all need to follow their instructions. Standard team reasoning has each individual doing the computation, identifying her own component, and following her own instructions. However, we can imagine another team mechanism, which Bacharach called *simple direction*, where the computing, identifying, and instructing is done by one individual—the *director*—and all the other team members need do is each follow the instruction that the director communicates to them. Provided that the team members have reason to think that the director's judgment has some validity, then each of them only needs to play their part. There is no need for computation or reasoning by every team member. (This may be similar to Ferrero 2010's idea of a "division of deliberative labor".)

The director is a role and how that role is filled may be determined in different ways. For example, in a football team, the coach may act as director. In this case, the director is a particular person, who permanently has a different role from the rest of the team. However, in other situations the role of director may be taken up by different team members at different times. For instance, take the collective discussed above, where different individuals take it in turns to lead and to make decisions, or an academic department, where the Head of Department role rotates among the members.

There are also different ways of communicating the plan. The football coach may announce the plan to everyone in the locker room; the Head of Department may announce the plan to everyone at a meeting. However, sometimes team members do not need to know the whole plan, knowing their part will suffice. The Head of Department may tell each member separately what they need to do. In a resistance cell, it may be a matter of safety that members only know their own part of the plan. Sometimes in drama productions the director may only tell each actor their own part, in order to produce a dramatic effect. Continuing the example of collective self governance from above, where the members take it in turn to be the leader, if the leader computes the plan and communicates the individuals' parts to them, and the individuals act on that plan, then the leader is the director and they are taking it in turns to be the director.

Analogously, in the intrapersonal case, we can think of the first timeslice as the director. That timeslice can compute the best team plan and commit it to memory, which is the way that the plan is communicated to later timeslices. This plan is an intention. If later timeslices identify with the person and remember their part in the plan, then all they need to do is play it. An intention is a part of a team mechanism for persons, which solves the problem that it is overly demanding to expect every timeslice to reason. In the language of decision theory, an intention is a solution for agents who are boundedly rational and do not have infinite time or capacity for computations. Any timeslice who finds themselves identifying with the person can simply retrieve their part in the plan from memory and carry it out.

That leads us to the second way that the simple account of team reasoning may be over demanding: It is implausible that every timeslice will identify with the person and team reason. Team mechanisms, including team reasoning, achieve a common interest or a common goal. (From now on, when I talk about team reasoning I will take it to encompass team mechanisms more broadly.) However, situations of temptation pit the interest of the timeslice against the interest of the whole. Again, we can turn to the interpersonal theory to find solutions. Bacharach (2006) thought there could be a local failure in some of the individuals in the team. He introduced the idea of *restricted team reasoning*, for situations in which it is known for certain that a team member or members will fail to team reason. In that case, members who team reason look for the best team plan, given the non-cooperative actions of the others.

Of course, team members may not be certain about whether other members will team reason. In the case of temptation, the same person may sometimes give in and sometimes not. In the interpersonal case, Bacharach formulated *circumspect team reasoning* in order to deal with this situation. He introduced a probability, omega, that individuals will group identify. Any individual who identifies and is in the position of computing a team plan can take this probability into account in their calculations. Circumspect team reasoning is particularly relevant to situations where there is some conflict of interest among team members. In the interpersonal case it can provide a model of cooperation in the prisoner's dilemma that admits the possibility that sometimes individuals will defect, either because they do not group identify or because they group identify, but judge that others are not likely enough to do so. Similarly, in the intrapersonal case, circumspect team reasoning can provide a model of temptation where timeslices sometimes but not always give in, and where earlier timeslices sometimes but not always form an intention to resist temptation, depending on how likely they think future timeslices are to resist (Gold, 2018). There can be circumspect direction by the timeslice that forms the intention.

This solves Puzzle 3: the earlier timeslice can rationally form the prior intention not to have a second biscuit, even if she know that, when the time comes to do as I intend, it will be rational for me qua timeslice not to follow through on that intention. The intention is a contingency plan, for later timeslices who end up self identifying.

So far, unlike in RNR accounts, the intention has had no causal role in helping overcome temptation: if the timeslice self identifies, then she follows the plan. However, there is a second role that intrapersonal planning can play: having the plan could prompt self identification. Standard decision theory allows earlier timeslices to take actions to constrain later timeslices. Analogous to this, 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 self identify. For instance, it makes salient the existence of the temporally extended person and the shared interests of the timeslices. Some researchers have argued that it is difficult to experience memories without also experiencing the sense of personal identity (Garfield, Strohminger & Nichols, 2018, and references within). By encouraging the later timeslice to identify with the person and therefore to act on the plan, the intention may prevent the transient-agent reasoning that leads to weakness of will. So it can have the same function as Holton's (1999, 2009) resolutions, a subset of intentions that have the specific purpose of overcoming contrary inclinations at the time of action. However, rather than stopping reconsideration, the resolution works by increasing the probability of self identification (which potentially also makes the timeslice less likely to engage in timeslice-level reasoning). The earlier timeslice can form the plan with the intention that it will play this causal role in resisting temptation. Nevertheless, in the model of intra-personal team reasoning, the resolution is not the root cause of self-control. Remembering the plan prompts team reasoning, so the effectiveness of resolutions is parasitic on the mechanism of intrapersonal team reasoning, which underpins self control.

6 | WHY DO INTRAPERSONAL TEAM REASONING AND RESIST TEMPTATION?

I have shown how it can be instrumentally rational to exert self control if a timeslice self identifies. However, it can also be instrumentally rational to give into temptation if the timeslice does not self identify. Many people have the intuition that we ought to exert self control. Can we say anything stronger and make a case that the timeslice ought to self identify with the person?

We will not be able to make any headway using instrumental rationality alone. Instrumental reasoning is about adopting suitable means to one's ends, taking one's ends as given. Instrumental rationality does not evaluate the ends themselves. Therefore a theory of the instrumentally rational choice of the unit of agency would require that one's ends were fixed before the choice of unit occurs. However, one's ends may be intimately bound up with one's unit of agency. The evaluation of an outcome may depend on the unit of agency with which one identifies (for instance the state where I eat two biscuits is evaluated differently from the perspectives of the timeslice and the person). Different units of agency are associated with different ends. For instrumental rationality, all goals are the goals of agents; there is no agent-neutral "view from nowhere" within instrumental rationality that can set one's ends. Therefore, choosing a unit of agency cannot be done independently from choosing ones ends. The unit of agency—and the concomitant evaluation of ends—must be fixed before the instrumental reasoning can occur.

This is a problem for accounts of "resolute choice", which maintain that being rational requires acting in accordance with the deliberative procedure that best serves one's concerns (McClennen, 1990; Gauthier, 1997). Why privilege the person over the timeslice or, indeed, wider humanity?[1] Indeed, this is a problem for any account that appeals to the consequences or benefits of identifying with a particular unit of agency. That includes psychologists with an inclination to intrapersonal Utilitarianism in their views of how to aggregate experiences. The way that people evaluate temporally extended outcomes may be different from the instant evaluations that they report in real time (Kahneman et al., 1993). Kahneman, Wakker, and Sarin (1997) recommend we take the instant reports of "experienced utility" and aggregate them, giving a measure of individual well-being that is neutral with respect to the time that experiences occur.[2] But why be neutral with respect to time but not with respect to persons?

The problem with consequentialist accounts is that, without a reason to privilege the personal-level, we are pushed towards a full consequentialist agent neutral morality, which is neutral between timeslices of different persons. So, for instance, Parfit (1984) argued that morality requires us to take into account the interests of our future selves, to the extent that they can be considered as a separate person. However, for exactly the same reasons, we can be obliged to take into account the interests of other people. Taking the consequentialist position seriously leads one to Hurley's (1989, p.136) position, where we should "contribute to the best unit of agency possible in the circumstances", and this may not be the individual.

The case for self identifying requires a reason for privileging one unit of evaluation over the other. Although the normative support will not come from instrumental rationality, there are several other possibilities for where this support could come from.

One possibility, which we can quickly dismiss, is that we privilege one of the perspectives because it is the decision maker. This is potentially appealing in the interpersonal case. When comparing individuals and groups, we might think that practical reasoning is done by individuals, who are the locus of decision making, and it is therefore natural to privilege the individual perspective. However, the real locus of decision making is the timeslice, it just happens that most of the time an individuals' preferences are stable over the timeframe being considered. Once we think of persons as timeslices, the case for privileging the evaluative stance of the locus of decision making looks less appealing. Indeed, in the case of temptation, privileging the timeslice is the opposite of most people's intuitions.

It is more promising to look outside the resources of decision theory. I will sketch two different routes by which we could go.

One rationale for privileging the person-level derives from our need for practical unity over our lifetimes. Many of our projects extend over time and some of our ultimate ends, like preserving our health, presuppose an ongoing identity. On this view, in order to be an agent, one has to perceive oneself as a persisting whole. It resonates with James' (1890) idea that we are a unity because we recognise that selves at different times are all a part of the same person (Gold & Kyratsous, 2017). There are a couple of different ways that we can cash this out. Schechtman (2007, 2008), in an account of the person that also allows both a temporally-limited and a temporally-extended perspective, takes an internalist position, where it is our belief that we have interests that extend over time that allows us to construct our personhood. However, Korsgaard (1989) takes an externalist view, where our practical projects give us a reason for regarding ourselves as being the same person as the self who will occupy our body in the future. The worth of achieving our projects grounds our unity as a person over time; our reasons extend over time, which drives us to a unity where our actions come from our whole selves. This aligns with a more traditional picture of akrasia, where resisting temptation is rational because that is what is required by the totality of one's reasons (Davidson, 1969). However, it is not clear that this Kantian rationale will always privilege the level of the person. One might wonder whether our needs to pursue our projects would sometimes drive us to identify with a group, a unit of agency that is higher-level than an individual (Anderson, 2001).

Finally, we have the approach of Bratman (2014, 2018), who attributes to the agent a non-instrumental end of self-governance. Bratman takes from Frankfurt (1987) the idea of a *standpoint,* a web of relevant attitudes that speaks for the agent and that guides her thoughts and actions. In order to play that role, the attitudes in the standpoint need to be sufficiently coherent and any synchronic standpoint will be "substantially plan-infused" because it will involve plans for temporally extended activities and general policies for action and deliberation (Bratman 2018, p.6). The non-instrumental concern with diachronic governance is an end within the agent's standpoint. Therefore, an agent who has a resolution but who re-opens the question at the time of temptation can do further reasoning in order to exert self-control. The present judgment conflicts with other elements of the standpoint, but the end of self-governance puts pressure on agents to ensure consistency. Therefore it can lead to a shift in the agent's standpoint, so that it is in favour of exercising willpower. For Bratman, anticipating the regret which would follow from giving into temptation can play a role in this reasoning, by shifting the agent's standpoint in a manner that favours sticking with the resolution. Bratman specifies that the non-instrumental end of self-governance is not rationally required, nevertheless there are good reasons to be a planning agent. Having a non-instrumental end of self-governance is a part of a stable rational equilibrium and a planning agent who reflects on the structures of her self-government can see this.

7 | CONCLUSION

I started out with three puzzles, thrown up by previous accounts of how we can use intentions to guard against temptations: How can it be rational not to give into temptation? How can the person, at the time of temptation, be motivated to follow a plan that is contrary to its current judgment? How can it be rational to form a plan to resist if we recognise the rational priority of the present judgment?

Intra-personal team reasoning (understood to include team mechanisms) can solve these puzzles: It is rational not to give into temptation if the timeslice identifies with the self over time. If the timeslice identifies, then it is motivated to follow the plan. Therefore it can be rational to form the intention to resist temptation as a contingency plan, for timeslices that identify. Further, a timeslice can form a resolution with the aim of prompting team reasoning later on, so the plan is not causally inert. However, in order to justify why the timeslice should do intra-personal team reasoning, we will need a thicker account than mere instrumental rationality. I have sketched two different forms that such an account could take.

Despite the major differences between Bratman's (2014, 2018) and Holton's (2009) accounts of how we guard against temptation, they are united in placing diachronic rationality at the centre of their accounts. Instead, in my account of intrapersonal team reasoning, I showed how adding a second level of agency allows a solution within the requirements of synchronic rationality, one which allows diachronic rationality to align.

ORCID

Natalie Gold https://orcid.org/0000-0003-0706-1618

NOTES

[1] In the interpersonal case, attempts to show that it is instrumentally rational to cooperate have failed. This is not the place to rehearse the debate, but note that even Gauthier (2013), assessing his research program since his original paper (Gauthier, 1986), admits there is a missing step in his argument.

[2] To be fair, the picture is actually a little more nuanced than this: later Kahneman claims that instant utility is not a purely hedonic concept (Kahneman, 1999) and that he does not propose this method as a comprehensive concept of wellbeing, only as a significant constituent of it (Kahneman, 2003).

REFERENCES

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person* (pp. 139–58). Cambridge: Cambridge University Press.

Anderson, E. (2001). Symposium on Amartya Sen's philosophy: 2 Unstrapping the straitjacket of'preference': a comment on Amartya Sen's contributions to philosophy and economics. *Economics and Philosophy*, *17*(1), 21–38.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American psychologist*, *37*(2), 122.

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. (2018). *Planning, time, and self-governance: essays in practical rationality*. Oxford University Press.

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

Bratman, Michael. (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge University Press.

Broome, J. (2013). *Rationality through reasoning*. John Wiley & Sons.

Davidson, D. (1969). How is weakness of the will possible?. in Joel Feinberg (ed.), *Moral Concepts*, Oxford: Oxford University Press.

Ferrero, L. (2010). Decisions, diachronic autonomy, and the division of deliberative labor. *Philosopher's Imprint*, *10*(2), 1–23.

Frankfurt, H. (1987). Identification and wholeheartedness. In Ferdinand David Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Gauthier, D. (1997). Resolute choice and rational deliberation: A critique and a defense. *Noûs*, *31*(1), 1–25.

Gauthier, D. (2013). Twenty-five on. *Ethics*, *123*(4), 601–624.

Garfield, J., Nichols, S. and Strohminger, N. (2018). Episodic Memory and Oneness. In P. Ivanhoe, H. Sarkissian, E. Schwitzgebel, O. Flanagan and V. Harrison(Eds.) *The Oneness Hypothesis: Beyond the Boundary of Self*. New York: Columbia University Press.

Gold, N. (2018). Putting Willpower into decision theory: The person as a team over time and intra-personal team reasoning. In J. Bermudez (ed) *Self-control and Rationality*. Cambridge: Cambridge University Press.

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

Gold, N. (2012). Team reasoning and cooperation. In S. Okasha and K. Binmore (Eds). *Evolution and Rationality: Decisions, Cooperation and Strategic Behaviour*. Cambridge: Cambridge University Press

Gold, N., & Colman, A. M. (2018). *Team reasoning and the rational choice of payoff-dominant outcomes in games.* Topoi, *39*, 305–316.

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, N., & Sugden, R. (2006). *Conclusion. In Bacharach, M. Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Gold, N., & Sugden, R. (2007). Theories of team agency. In F. Peter, & H. B. Schmid (Eds.). *Rationality and commitment*. Oxford University Press.

Heyman, G. M. (2009). *Addiction: A disorder of choice*. Harvard University Press.

Hollis, M. (1998). *Trust within Reason*. Cambridge University Press.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press.

Holton, R. (2004). Rational resolve. *The Philosophical Review*, *113*(4), 507–535.

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

Hurley, S. L. (1989). *Natural reasons: Personality and polity*. Oxford: Oxford University Press.

James, W. (1890). *The principles of psychology*. New York: Holt.

Kahneman, D. (1999). Objective happiness. *Well-being: The foundations of hedonic psychology*, *3*(25), 1–23.

Kahneman, D. (2003). Experienced utility and objective happiness: A moment-based approach. In D. Kahneman, & A. Tversky (Eds.) *The psychology of economic decisions*, *1*, 187–208.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological science*, *4*(6), 401–405.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The quarterly journal of economics*, *112*(2), 375–406.

Karpus, J., & Gold, N. (2016). Theory and evidence. *The Routledge Handbook of Philosophy of the Social Mind*, 400–17.

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

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In A. E. Roth, & J. H. Kagel *The handbook of experimental economics* (Vol. 1). Princeton, NJ: Princeton University Press.

McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge university press.

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

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

Paul, S. (2011). Review of Richard Holton, Willing, Wanting, *Waiting. Mind. 120*, 889–892

Pettigrew, R. (2020). *Choosing for changing selves*. USA: Oxford University Press.

Pettigrew (2019). Review of José Luis Bermúdez (ed.), *Self-Control, Decision Theory, and Rationality: New Essays. NDPR* https://ndpr.nd.edu/news/self-control-decision-theory-and-rationality-new-essays/Accessed Oct 15 2019.

Read, D. (2001). Intrapersonal dilemmas. *Human Relations*, *54*(8), 1093–1117.

Schechtman, M. (2008). Diversity in unity: practical unity and personal boundaries. *Synthese*, *162*(3), 405–423.

Schechtman, M. (2007). Stories, lives, and basic survival: A refinement and defense of the narrative view. *Royal Institute of philosophy supplement*, *60*, 155–178.

Schelling, T. C.(1978). *Micromotives and Macrobehavior*. Norton & Co: New York.

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

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 165–180.

Thoma, J. (2018) Temptation and preference-based instrumental rationality. In José Bermudez (ed.), *Self-Control, Decision Theory, and Rationality*, Cambridge University Press.

How to cite this article: Gold N. Guard against temptation: Intrapersonal team reasoning and the role of intentions in exercising willpower. *Noûs*.2021;1–16. https://doi.org/10.1111/nous.12369