Individual and Social Commitment

John Michael
johnmichael.cogsci@gmail.com

Draft – Comments Welcome

Abstract

This paper examines the relationship between individual and social forms of commitment. It spells out a framework showing how social commitment builds upon individual commitment. The foundation of this framework is an analysis of the core function which unites many, though probably not all, instances of commitment: namely, to shield goals that are in one's long-term interest from fluctuations in short-term interests and impulses. This functional analysis makes it possible to discern the underlying cognitive and motivational mechanisms which are common to instances of individual and social commitment, and to develop a comprehensive overview of individual and social factors that may trigger or enhance commitment. Along the way, testable predictions are generated by the framework, existing evidence is reviewed and re-assessed in light of the analysis offered here, and a range of new questions are articulated for further theoretical and empirical research.

Keywords: Commitment, goal shielding, self-control, sunk cost reasoning, cooperation

1. Introduction

In everyday life, we experience many different forms of commitment. Consider the following four examples:

a) Agnes made a commitment to pick Sam up at the airport tomorrow.
b) Polly and Pam are in the habit of smoking a cigarette and talking together on the balcony during their afternoon coffee break. They have never explicitly agreed to do this, but Polly is aware that Pam expects her to show up today, like every other day.
c) Frank was unsure whether to go to the cinema or the theater tonight, but he decided in favor of the cinema and now he is committed to that plan.
d) Roger is committed to birdwatching and spends considerable amounts of time and money pursuing this hobby.

There are many differences among these four examples, and they could be used to illustrate a number of distinctions which one might make among different categories of commitment. For example, sometimes commitments pertain to specific one-off goals to be met (Agnes and Frank), whereas sometimes they imply an ongoing level of interest or a determination to persist in the face of obstacles (Polly and Roger). Sometimes they involve obligations (Agnes and possibly Polly), whereas sometimes they do not (Frank and Roger)[1]. One very general distinction which these examples illustrate is the distinction between social and individual commitment: Agnes and Polly are committed at least in part because the goals in question are ones that are valuable to other people, whereas this is not the case for Frank and Roger. This distinction will be my focus here.

Specifically, the current paper examines the relationship between instances of individual commitment and instances of social commitment. Of course, we should not assume that they have anything interesting in common simply because we sometimes use the same English word to refer to them. However, as we shall see, both categories

[1] There are many other differences as well among these cases and other cases of commitment. For attempts to taxonomize heterogeneous cases of commitment, see Löhr (in prep); Michael & Pacherie (2015); Shpall, 2014.

can be illuminated by careful consideration of the ways in which they relate to each other. In particular, we shall see that it is fruitful to think of social commitment as building upon individual commitment.

I will begin by reviewing two contrasting approaches to commitment, each of which provides useful insights to be incorporated into a comprehensive framework. The first of these approaches, drawing on Bratman (1987; 2017), takes individual commitment as the starting point: it attempts to conceptualize individual commitment and to use this as a basis for explaining social commitment. The second, a game-theoretic approach, takes social commitment as the starting point (Frank, 1988; Schelling, 1980). This approach invites us to view individual commitment as a derivative of social commitment (Michael & Pacherie, 2015; Heintz, in prep). I evaluate these approaches on their own terms, identifying strengths and weaknesses of each (sections 2 and 3), and then turn to the development of a comprehensive framework (section 4). The foundation of this framework is an analysis of the core function which unites many, though probably not all, instances of commitment. This functional analysis will also enable us to discern the underlying cognitive and motivational mechanisms which are common to instances of individual and social commitment, and to develop a comprehensive overview of individual and social factors that may trigger commitment. Along the way, we specify testable predictions generated by the framework, review existing evidence bearing upon those predictions, and articulate a range of new questions for further theoretical and empirical research.

2. Individual Commitment: Bratman

Bratman's starting point is to examine the role of intentions in individual agency. On his analysis, intentions function to terminate practical reasoning and to structure means-end reasoning about how to achieve goals. In other words, they settle the question of what goal to pursue, and thereby enable one to move on to the subsequent question of how to go about achieving the goal. Taking one of our examples from the introduction: Frank desires equally to go to the cinema and to the theater, but cannot do both because the performances are at the same time, so he finds it difficult to form a plan to do either But if he forces himself to make a decision in favour of the one or the other, he forms

an intention to do the one or the other. Now he can end his deliberations and use this intention as a basis for forming a plan.[2]

In order for the intention to fulfill these functions, it has to have at least some degree of robustness: if Frank decides to go to the cinema but then, when confronted with the need to decide which metro line to take (assuming he would need to take the blue line to get to the cinema and the green line to get to the theater), he again starts deliberating about whether he prefers the cinema or the theater, his original decision and his resultant intention will not really have served their purpose. In other words, intentions are helpful in part because they involve commitment to a course of action.

However, it would also be silly to stick blindly with intentions in the face of important new information. If it turns out that the metro line running to the cinema is under construction and Frank would have to walk, then maybe it makes sense for him to reconsider. In other words, intentions should not commit us *unconditionally*. As Castro & Pacherie (2020, p.9) have recently pointed out, this means that intentions actually require us to perform a balancing act between pusillanimity and stubbornness. One limitation of Bratman's account is that it does not specify any normative principles which would help to determine when it is rational to persist and when it is not.

A second limitation of Bratman's account is that it does not illuminate the underpinning psychological mechanisms which determine whether and to what extent we remain committed to our intentions, nor the mechanisms which then actually do sustain commitment. It must be acknowledged of course that Bratman's account is not designed to illuminate these mechanisms. Be that as it may, if we are interested in understanding the psychological mechanisms underpinning commitment, we will have to look elsewhere.

What about social commitment then? Bratman (1987) points out that social context may bolster the case for resisting reconsideration in cases in which we have stated our intentions publicly, because we may want to maintain our reputation as predictable, reliable agents so that others will be willing to interact with us in the future (Theriault, Young, & Barrett, 2020). This implies that if others are aware of our intentions, we should for this reason be more resistant to reconsideration than we otherwise would be. This points to a third limitation of Bratman's account. Specifically, Bratman makes his

[2] Having to evaluate alternatives and make choices in such cases can be debilitating for patients with frontal lesions (Damasio, 1994).

point about the social dimension of commitment in relation to cases in which an intention has been publicly stated. But why should this be decisive? Roger's motivation to go birdwatching may be enhanced if other people are aware of his commitment to birdwatching, but this does not require him to have stated it publicly; it may be sufficient for other people to have observed him birdwatching regularly or to have heard him talking about birds. Thus, Bratman is right that we may bolster our individual commitments by drawing other people into them. But stating them publicly is just one way of doing this, not a necessary condition. This raises the question under what circumstances other people's expectations or reliance may bolster our individual commitments

In sum, Bratman's analysis provides a clear and compelling reason for thinking that goal-directed action in general requires a certain degree of commitment. This is because we need to settle some practical questions (e.g. what goal to pursue) in order to get on to other questions (What plan to pursue? What goal to aim for next, etc). He also provides the starting point for an analysis of how social commitments can build on individual commitments: our relationships with others and our reputations may be affected by the amount of commitment we exhibit. Building on Bratman's analysis, it would be desirable to identify normative principles bearing upon the question as to how much commitment is appropriate, and under what circumstances more or less commitment is appropriate. It would also be valuable to develop a better understanding of the psychological mechanisms by which we determine how much commitment is appropriate and by which we implement that level of commitment. Finally – and of particular interest to us here – we would like to identify the social factors that may build upon and bolster individual commitment.

3. Social Commitment: A Game-Theoretic Perspective

In the context of game theory, a commitment is a particular way of solving strategic problems where one agent would like to get a second agent to do something, but where that second agent is only willing to do so if the first agent agrees to do something else (or to refrain from doing something else). The challenge for the first agent is to persuade the second agent that s/he (the first agent) really will do as s/he says at that later point

in time, after the second agent has lost her leverage. Generally speaking, commitment in this game-theoretic context can be thought of as a 'device to leave the last clear chance to decide the outcome with the other party, in a manner that he fully appreciates; it is to relinquish further initiative, having rigged the incentives so that the other party must choose in one's favor' (Schelling, 1980: 37). For example, one might attempt to win a game of chicken (a.k. hawk-dove) by removing the steering wheel and holding it out the window for the other driver to see, and thereby removing the option of swerving to avoid the other driver. This strategy effectively forces the other driver to make the final decision whether to collide or to swerve.

One common means of solving this type of problem is by signing contracts. The way contracts solve this type of problem is by changing the parties' incentives: if the first agent does not do what s/he has promised as the later point in time, s/he faces a penalty – and the second agent, knowing that this is the case, is thereby assured that the first agent will in fact do as she says. Often, however, contracts are not possible. To borrow one of Schelling's (1980: 43-44) examples, a hostage would like to persuade his captor to release him, which the captor is in principle willing to do (e.g. because it has become clear that the ransom will not be paid). The captor may hesitate out of fear that the hostage will testify against him. Of course the hostage could promise not to, but why should the captor believe this promise? To solve this problem, what the hostage needs to do is to somehow change his own incentive structure such that testifying against the captor is not in his own interests. One way to do this is to confess to some other crime (or to commit some other crime), and to provide the captor with evidence of this which the captor could use against one.

A more commonplace strategy, when formal contracts are not feasible, is for one agent to put her reputation at stake by making her commitment to a second agent public (Michael & Pacherie, 2015; Heintz, Karabegovic, & Molnar, 2016). This way, if she doesn't perform the action that she committed to, she may suffer reputational costs. Down the line of course, this is likely to have material implications insofar as she may find it more difficult to find partners for mutually beneficial cooperative endeavors, as illuminated by partner choice models of mutualistic cooperation (Barclay & Willer, 2007; Baumard, André & Sperber, 2013). Thus, it may be in the first agent's material interest to act in accordance with her commitment even in the absence of a formal contract.

This game-theoretic perspective provides a clear functional task description for commitment: a commitment is a deliberate and discrete act by which an agent changes the payoff structure of her own future options in order to convince some other agent that she will do one thing and not another at a future time point. It is worth highlighting that commitment in this sense is *strategic*: the first agent commits because she wants to convince the second agent that she will do something later on in order to get the second agent to do something else first (or at the same time).

It is worth highlighting three important features of this approach, because these features will provide useful points of contrast with the framework that we will be developing here:

First, the game-theoretic approach takes social instances of commitment as its starting point – i.e. as opposed to instances of individual commitment, such as when one is committed to birdwatching or to breaking the hot dog eating record. Does this mean that it cannot be applied to instances of individual commitment? Not necessarily. One possibility is to consider that we can deliberately draw in other people in order to change the incentive structure of our own individual actions. For example, in order to enforce one's intention to maintain a diet, one can enter into a wager with a third party and thereby introduce an extra penalty for non-compliance (Luce & Raiffe, 1989). A further possibility is to think of cases of individual commitment as cases in which one makes a commitment to one's future self (Sperber & Baumard, 2012). In this sense, failing to get up early enough to go birdwatching may lead Roger to disappoint his own expectations about himself. This would imply that the aversion to disappointing oneself in the future provides an additional motivation to act in accordance with one's commitment, lowering the net value of alternative options (i.e. skipping a day of birdwatching to watch TV series). This conjecture gains face value from research showing that when people feel more strongly connected with their future selves, they tend to be more willing to forego current rewards to obtain larger rewards later in time (Bartels & Rips, 2010). But before attempting a thorough evaluation of this extension of the game-theoretic concept of commitment to individual commitment, it would be important to fill in further details. For example, we should ask whether there is some equivalent of the deliberate act by which a committing agent changes her future payoff structure, and also whether there is some equivalent of the strategic function of persuading some other agent to do something that she would not otherwise do. Often

there seems not to be either of these things. How then do individual and social commitment relate to each other? We will return to this below.

A second feature we would like to highlight is that the game-theoretic account is tailored to cases in which a commitment is generated deliberately and explicitly. As a result, it does not illuminate the conditions under which commitments can arise unintentionally or gradually. To illustrate, consider the example of Polly and Pam cited in the introduction to this paper (also described by Michael, Sebanz & Knoblich, 2016: 3; adapted from Gilbert, 2009: 6; for evidence that many people share the authors' intuitions about this example, see Bonalumi, Isella & Michael, 2019):

> Polly and Pam, are in the habit of smoking a cigarette and talking together on the balcony during their afternoon coffee break. The sequence is broken when one day Pam waits for Polly but she doesn't turn up. In this case, there has been no explicit agreement to smoke a cigarette and talk together every day, and yet one might nevertheless have the sense that an implicit commitment is in place, and that Polly has violated that implicit commitment. This will depend on further details about the case. For example, if Polly and Pam have smoked and talked together every day for 2 or 3 weeks, Polly might feel only slightly obligated to offer an explanation, but she would likely feel more strongly obligated if the pattern had been repeated for 2 or 3 years. Thus, it seems that mere repetition can give rise to an implicit sense of commitment. Similarly, […] one agent's investment of effort or other costs in a joint action may also give rise to an implicit sense of commitment on the part of a second agent. If Pam, for example, must walk up five flights of stairs to reach the balcony where she and Polly habitually smoke together, Polly's implicit sense of commitment may be greater than if Pam only had to walk down the hall.

Is it possible to map the game-theoretic conception of commitment onto these cases? There seems to have been no act by which Polly changed her payoff structure to prop up the value of the option of going to the balcony. One possibility is to think of it as the series of actions of going out onto the balcony rather than any single discrete action. But why should the repetition of the action (or Pam's investment of effort costs) increase Polly's valuation of the option of going to the balcony (or decrease the valuation of alternative options)? Some explanation would need to be provided of why these factors make a difference with respect to the reputational costs incurred by Polly if she does not show up. It is also worth highlighting that in this scenario Polly has not performed the action previously with the strategic intention of persuading Pam that she would do it in the future.

A third feature of this game-theoretic approach to commitment which we would like to highlight is that it conceptualizes commitment as an act with a particular pragmatic function, not as a psychological phenomenon. As a result, it is neutral with respect to the cognitive and motivational processes enabling individuals to commit or to remain committed – i.e. to resist short-term temptations and to act in accordance with commitments which optimize their long-term interests. From a normative point of view, it may be tempting to brush this psychological level aside. But it is well-known that people are often tempted to make myopic decisions which fail to maximize their long-term benefits, and that they often succumb to such temptations (for an overview, see e.g. Read, 2004). Why is it often tempting not to do what is in one's best interests in the long-term? How do people often manage to resist such temptations?

We started out this section by characterizing commitment in game-theoretic terms. This provided us with a clear conception of commitment as a particular kind of deliberate act that serves a particular strategic function. We also identified three distinctive features of this approach to commitment: it takes social rather than individual commitment as its starting point, it is clearly tailored to cases of explicit rather than implicit commitment, and it is neutral with respect to the cognitive and motivational mechanisms underpinning commitment. Neither of these is a problem per se, but it would be desirable to develop an account which related individual to social commitment, which specifies the circumstances giving rise to implicit commitment (as well as the factors which modulate the degree of commitment, as the example of Polly and Pam also illustrated), and which illuminates the cognitive and motivational mechanisms underpinning commitment.

4. A Comprehensive Framework

4.1 Limitations of Motivational Integration

In constructing a framework that illuminates the deep relationship between individual and social commitment, my starting point is the assumption that we often are tempted to act in ways that do not support our long-term interests. For example, one may be tempted to stay in bed and sleep for a few extra hours rather than going to work, or to smoke a cigarette, eat a second piece of cake or drink an extra glass of wine while out at a party. Everyday examples like these reveal the limitations of our motivational

integration – i.e. they illustrate that our currently predominant motivation does not always adequately reflect what is our long-term interests.

This apparent limitation of motivational integration is in fact supported by decades of research on reward processing and motivation. This research reveals that the regulation of motivation results from a complex interplay of distinct mechanisms which track and respond to different, imperfectly aligned indicators of value. One central distinction is that between mechanism for 'liking' and mechanisms for 'wanting'. "'Liking' is essentially hedonic impact—the brain reaction underlying sensory pleasure-triggered by immediate receipt of reward such as a sweet taste….'Wanting', or incentive salience, is the motivational incentive value of the same reward...'Wanting' is purely the incentive motivational value of a stimulus, not its hedonic impact" (Berridge, 2004:194). Liking and wanting normally go together: you want to eat when you are hungry (food has incentive value that motivates you to eat), and you like the experience of eating (you experience pleasure while eating). But they can come apart as a result of certain experimental manipulations and in certain pathological cases. In addiction, for example, people may want a drug but not take pleasure in it – i.e. not like the actual experience (Robinson & Berridge, 1993; 2003).

This observation of limited motivational integration provides us with the core of a functional task description for commitment: to shield goals which are in our long-term interests from fluctuations in our short-term interests or current impulses. Given this functional task description, it is apparent that there may be many reasons why a particular goal is in our long-term interests – but these differences in the source of long-term value of goals do not entail that the mechanisms which shield goals from fluctuations in short-term interests or current impulses need to be different (Shah, Friedman & Kruglanski, 2002; Dreisbach & Haider, 2009; Hofmann, Friese, & Roefs, 2009).

What are these mechanisms? Hofmann, Friese, & Roefs (2009) distinguish three ways in which goals may be shielded: (i) attentional control is engaged to exclude information which is not relevant to the pursuit of the goal from being noticed; (ii) by exercising inhibitory control to avoid performing actions or entertaining thoughts which are not conducive to the goal; (iii) affective control to enhance positive emotions arising from goal pursuit and to dampen negative emotions arising from resistance to temptation and distraction. These three types of goal shielding may work in concert or

independently of each other – and indeed, if attentional and/or affective control are sufficiently effective, the demands on inhibitory control may be reduced.

Thus, cases of individual and social commitment can be seen to differ with respect to the source of the goal or the reason why the goal is in our long-term interests, but not necessarily with respect to the mechanisms engaged in stabilizing our motivation to act towards the goal. This already provides us with a partial answer to the question how individual and social cases of commitment relate to each other: they are likely to overlap substantially with respect to the mechanisms which they engage to stabilize motivation to act towards valuable goals, and they are likely to differ with respect to the sources of the goals in question.

Taking a step further, we also wish to understand how those aspects of individual and social commitment which differ (i.e. the sources of goal valuation) relate to each other. To do this, we will need to examine how goals come to be identified as being in our long-term interests, and in particular to home in on the individual and social factors which lead us to identify some goals as being in our long-term interests, and thus worth shielding from fluctuations in short-term interests and current impulses.

4.2 Goal selection and progressive valuation

In order to structure our examination of how some goals come to be identified as being in our long-term interests, it will be helpful to think in terms of a progression across various stages: from selecting a goal, to forming a plan, initiating action in pursuit of the goal, and persevering all the way through to completion. To be clear: this full template for goal progression is not applicable in all cases. We may sometimes find ourselves acting in pursuit of a goal without having deliberate selected it or engaged in any conscious planning. Nevertheless, the full template will enable us to identify (individual and social) factors which can come into play to increase goal valuation along the way through the progression – e.g. as a result of having selected, planned, initiated action towards the goal, etc. To identify and catalogue these factors, we will take one of the simple examples from the introduction and embellish it as we go along in order to distinguish among a range of possible cases in which these different factors come into play at various stages of goal progression.

Case 1: Goal Selection

Roger is a birdwatcher and is out hiking in the woods. He catches a glimpse of the characteristic twinkle on the wing of a slender-billed curlew, and is inclined to chase after it. Just then, however, it occurs to him that he has lost track of where he is, and he notices that it is getting dark and chilly. He realizes that it is in his best interests to give up on birdwatching and to build a shelter to sleep in.

Case 1 illustrates the need to select goals that are in one's long-term interests in the first place. We can think of commitment coming into play in this case in the sense that Roger has a stronger commitment to staying alive than to observing this particular bird. More generally speaking, he *values* some goals more than others in the first place. There are many reasons why some goals are particularly valued over others, e.g. because of a pre-existing valuation of some principle, some activity, a person, a relationship, or whatever. In such cases, it is not uncommon in everyday speech to use the term 'commitment' refer to this pre-existing valuation (e.g. being committed to a principle, an activity, a person or a relationship, etc.). We can think of commitment in these cases as a disposition to be committed to specific goals which are consistent with or serve the interests of the principle, activity, person, relationship, etc. to which we are committed.

What social sources might there be of commitment in this sense of goal valuation? It is useful to distinguish between two background functional hypotheses concerning possible social sources of commitment as valuation. First, one may *strategically* commit to others' goals in the present in order to increase one's likely future benefits – either through eliciting reciprocal prosocial behaviour from them in the future (i.e. direct reciprocity; Trivers, 1971) or through a boost to one's reputation (i.e. indirect reciprocity; Nowak & Sigmund, 1998). Secondly, Roberts' (2005) '*interdependence hypothesis'* could explain why one might be genuinely interested in the well-being of other group members. The interdependence hypothesis holds that humans' tendency to cooperate arose evolutionarily in a period in which our ancestors lived in small groups of individuals whose interests were largely interdependent, and for whom it was therefore not typically beneficial to act selfishly to the detriment of other group members. This implies that if a goal is valuable to some agent with whom one is interdependent, one should value the goal as well. As a result, any indication that one shares a valuable relationship with some other agent (that we are interdependent) and that a goal, G, is valuable to them should lead one to value G. Crucially, this line of reasoning does not depend on the expectation of reciprocity. This means that one

may value G because G is in some other agent's interest even though that agent does not know this, and may mistakenly believe that some other goal would be better for her. The interdependence hypothesis therefore predicts that one should in some instances be committed to goals on behalf of other people even though it may damage one's relationship with those other people and damage one's reputation.

It is worth noting that the strategic hypothesis and the interdependence hypothesis are not mutually exclusive, and also that they may overlap with respect to the proximal mechanisms with which they would predict evolution to have equipped us. For example, they are consistent with the idea that we have a preference for meeting others' expectations or of doing what others are relying on us to do (Dana et al., 2006; Heintz et al., 2015; Székely & Michael, 2018). According to the interdependence hypothesis, an agent expecting or relying on one to do G would be a cue that one has a relationship with that agent and that G is valuable to that agent. According to the strategic commitment hypothesis, meeting expectations and doing what others rely on us to do is an effective way to maintain working relationships and to manage one's reputation. But the two functional explanations would come apart in some cases, such as cases of the type described above – i.e. where one values a goal G because it is in some other agent's interests even though this other agent does not understand this and is likely to be angered. The interdependence hypothesis, but not the strategic commitment hypothesis, predicts that one should value such goals.

In sum, it is often in one's long-term interests to get along with others, to maintain relationships and to manage one's reputation. These long-term interests can be played off against short-term interests in just the same way as any other long-term interests. And insofar as we are committed to the broader goal of maintaining a relationship and/or another person's well-being, this commitment will regulate our motivations to ensure that we continue acting in accordance with it, adopting goals that contribute to maintaining the relationship and/or to furthering the person's well-being.

Case 2: *Planning*

As in case 1, Roger has decided to build the shelter and formed a plan to collect some large branches to make a frame and a bunch of spruce boughs for the walls and roof. He is just about to set to work. But now it occurs to him that he has a bunch of clothing which could conceivably be tied together to form a makeshift tarp, which could perhaps somehow be used instead of the spruce boughs…or maybe he doesn't even need the

frame if he just hangs everything from some trees…but will this really work? Instead of wasting time and energy evaluating this and the myriad other options that may occur to him, it may be wise just to stop thinking about it and get back to work.

In case 2, in contrast to case 1, it is not a question of selecting the appropriate goal in the first place but of resisting the temptation to reconsider what the most appropriate goal is. This is the kind of case which Bratman has in mind when he suggests that it is sometimes rational to resist reconsidering our options once we have settled on a plan. As noted above, in section 2, Bratman does not offer any normative principles for determining when and to what extent it is rational to resist reconsideration. One principle which on first blush seems compelling is that, when confronted with new information, one should consider whether one would have made a different decision if the new information had been available when one made the original decision in the first place. The problem with this principle is that to determine whether one would have decided differently, one needs to reconsider. In other words, the proposed principle does not really explain when one should reconsider.

A different approach is to rely on heuristics based upon a preliminary assessment of situation in light of the new information (i.e. without re-hashing the original decision-making process). What sort of heuristics might be most useful? In addressing this question, it is useful to bear in mind that re-consideration involves costs of various kinds.

First, there are opportunity costs insofar as the time spent re-considering could be spent implementing the current plan. These costs will be especially high in situations in which there is time pressure. This means that the *threshold for reconsidering an option one has selected should be higher when there is time pressure.* A prediction that follows from this is that people should be less likely to consider alternatives to the extent that they are under time pressure.

Moreover, the costs of re-consideration will depend on what one is doing already (to what extent the current course of action requires one's ongoing attention). A quick glimpse at alternative options may or not be possible without interfering with the current course of action. This means that *the threshold for reconsidering an option one has selected should be higher to the extent that the current action requires attention*. A prediction that follows from this is that people should be less likely to consider alternatives at all if the course of action they have chosen currently requires their attention.

It is also relevant that the costs and benefits of reconsidering options may be more or less certain. In case 2, it seems that both the costs and the benefits of reconsidering are relatively uncertain. It is not obvious whether or not the alternative plan is better than the one Roger is already implementing. To determine this, Roger would need to look around a bit to evaluate the available resources and imagine going through the steps of the alternative plan. And since the alternative option begins as just a vague idea, Roger won't really know right away how much time and effort it would take to work out the details and thoroughly evaluate the idea. In other cases, however, an alternative option may appear clearly right away. For example, if Roger were to notice an apparently abandoned hut among the trees of a nearby hillside, it would not require time-consuming deliberation to determine whether to change plans: the advantages of sleeping in the hut rather than building a shelter from scratch are immediately obvious. This contrast illustrates the point that *the threshold for reconsidering an option one has selected should be higher when the costs and benefits of doing so are uncertain*. A prediction that follows from this is that people should be less likely to consider alternatives to the extent that they take the decision landscape to be uncertain or volatile.

A related point is that one may have been more or less confident in the original plan in the first place. If Roger was not really sure to begin with that the shelter he was constructing would be sufficiently warm and dry, he should be more open to considering alternatives that arise than if he had been fairly confident in his plan. This means that *the threshold for reconsidering an option one has selected should be higher when one has high confidence in the original option*. A prediction that follows from this is that people should be less likely to consider alternatives to the extent that they were confident when making the original decision.

Finally, it is also relevant to consider how many other plans are likely to depend on the current goal being achieved, and would therefore also need to be abandoned or revised. In other words, *the threshold for reconsidering an option one has selected should be higher to the extent* that one has built on this goal in making further plans. A prediction that follows from this is that people should be less likely to consider new information to the extent that they have made further decisions or plans based upon the one that is currently a candidate for re-consideration.

This last point is particularly relevant for our discussion insofar as it provides a basis on which to build social commitment. This is because other people can also make

decisions and plans based on the expectation that one will achieve a particular goal. To the extent that other people may be aware that one has selected a particular goal, they may have made decisions or plans based on the expectation that one will act in pursuance of this goal. This means that if one does not follow through with the plan, others may be disappointed and may waste time or other resources – an outcome which one may prefer to avoid in general in order to preserve one's relationships and one's reputation (Dana et al., 2006; Heintz et al., 2015; Székely & Michael, 2018)[3]. A prediction that follows from this is that people should be less likely to consider new information to the extent that others are aware of the original decision they have made, and all the more so to the extent that others are likely to have made decisions or plans based on this expectation.

Case 3: Goal Pursuit

Roger has been working on the shelter for a while and gotten most of the frame up. Now he notices some other branches that may work even better, and they are in fact shaped just right so that he could build a frame out of them fairly quickly and would be done just as quickly as if he continued with the frame he has been working on so far. If he had seen these in the first place, he surely would have selected them rather than the ones that he did select. But now he thinks that it would be a shame to waste the effort he has already invested in the current frame.

This is sunk cost reasoning – i.e., Roger is taking his past investment into account in deciding how to act in the future (Heath, 1995). There is some controversy as to whether sunk cost reasoning is ever rational (Kelly, 2004; Walton, 2002). Though we need not address this controversy here, one argument in defense of sunk cost reasoning is relevant for the current discussion. Specifically, a tendency to take sunk costs into account may be useful as a heuristic that functions to keep one on track when one should stay on track but might be at risk of deviating. When might this be the case?

i) When one in fact should stay on track but is likely to form the mistaken belief that it is in one's interest to switch plans;

[3] Indeed, it can be argued one has a moral obligation to avoid disappointing the expectations which one has led others to form about one's future actions expectations, in particular when others are likely to be relying on those expectations – at least when those expectations are reasonable (Scanlon, 1998).

ii) When one is tempted for the wrong reasons to switch, e.g. because it is boring or effortful to continue, and one therefore comes to experience the task as aversive (Botvinick & Braver, 2015);
iii) When there is uncertainty about whether staying the course is the right choice (see case 2 above), and one is likely to waste time and energy if one starts reconsidering.

Interestingly, Heath (1995) argues that people don't engage in sunk cost reasoning very often except under very special circumstances, namely when it is difficult to calculate or compare the costs and benefits (see (i) and (ii) above). This would be consistent with the heuristic, identified above, that we should avoid re-consideration to the extent that the costs and benefits of alternative options are uncertain.

Sunk cost reasoning can be distinguished from what has been called soft commitment (Rachlin, 2016; Siegel & Rachlin, 1995). In soft commitment, merely beginning a behavioral pattern may increase the value of completing it, and as one progresses toward completion, the value of completion increases. To borrow an example from Rachlin (2016), a group of people playing baseball are going to be more willing to keep on playing despite a bit of rain if they have already reached the ninth inning than they would be if they had just started. Fictional examples aside, there has in fact been empirical research documenting soft commitment in humans (Kivetz et al., 2006), rats (Hull, 1932) and pigeons (Siegel & Rachlin (1995)[4]. Soft commitment is different from sunk cost reasoning because the whole pattern may be rewarding rather than costly. But like sunk cost reasoning, it implies a kind of 'mission creep' – i.e., the value of a goal is increased by acting towards that goal. We

Why on earth would acting towards a goal increase one's valuation of the goal? Why does having played eight innings make it more attractive to keep going until the end? One possibility (which would apply to soft commitment as well as to sunk cost reasoning) is that one's prior actions (selecting a goal, planning and initiating goal pursuit) may indicate to oneself that one values the activity and/or the goal (Schrift & Parker; Kivetz et al., 2006). A further hypothesis (which would apply to soft

[4] One way to distinguish experimentally between soft commitment and sunk cost reasoning is to control for the amount (of money, effort, time etc.) that has previously been invested. Soft commitment, in contrast to sunk cost reasoning, is sensitive to the distance *to the goal*, not the distance *from the starting point*.

commitment as well as to sunk cost reasoning) is that when one begins acting towards a goal, one tends to form other plans that presuppose the completion of that goal. When confronted with the option of abandoning the original goal, a sensible default procedure would be to consider whether any other plans would also thereby be affected. This leads to the prediction that people should resist reconsideration to extent that it is uncertain what other plans might be affected.

Commitment as mission creep provides a robust platform for social commitment: by initiating action towards a goal, one may lead others to form the expectation that one will complete the goal one has begun to pursue. This leads to the prediction that people should resist reconsideration to extent that others have observed them performing actions in pursuit of a goal, or indeed any action which are likely to be interpreted as indicating pursuit of a goal.

Indeed, this line of thinking suggests that by investing time, effort or other resources in persisting towards a goal, one indicates to others all the more clearly that one values the goal and will persist until one achieves it. Insofar as this may strengthen their expectation that one will achieve the goal, it also constitutes an invitation to them to rely on the expectation and to plan accordingly, thereby further entrenching one's own commitment towards to the goal, as well as theirs, etc. This conjecture gains credence from some recent research showing that the perception of a partner's investment of effort in a joint action increases people's sense of commitment to the joint action (Chennells & Michael, 2018; Székely & Michael, 2018).

From this perspective, making a promise or simply stating one's intention to perform a particular action appears as a special case of a broader class of cases in which, by initiating a sequence of actions which typically leads to a particular outcome, one invites others to form the expectation that one will bring about that outcome. In other words, the promise or the statement of an intention can be seen as a conventionalized first step in the sequence leading to a particular outcome. Of course, the act of making a promise introduces norms and obligations that may otherwise be absent. The claim here is not that promising to do X is no different at all from simply starting to do X, but it can be seen as building upon this broader phenomenon of creating expectations by initiating and persisting in goal-directed action. And indeed, Bonalumi, Michael, & Heintz (Under Review) have recently shown that a sense of commitment can be elicited 'if one agent (the sender) has led a second agent (the recipient) to rely on her to do something, and if this is part of the two agents' common ground. Crucially, this

situation can occur even if the sender has neither uttered a commissive speech act nor performed any action that would conventionally be interpreted as such.'

This line of thought also implies that there may be cases in which one accidentally creates expectations and (arguably) thereby generates a commitment unintentionally. Although I am not aware of any empirical evidence to support the hypothesis that this is the case, its face value is illustrated by an example taken from Michael, Sebanz & Knoblich (2016):

> 'Sam is cleaning up the living room and picks up a ball that had been lying on the floor. As it happens, his dog Woofer notices this and bounds over to him, apparently ready to play fetch. Sam was not intending to play fetch and does not particularly desire to, but may now feel obliged to, because he has generated an expectation on the part of Woofer that they will now play fetch together. Thus the unintentional generation of expectations can lead individuals to sense that a commitment is in place' (p.5)

5. Conclusion

In constructing a framework to illuminate the ways in which social commitment builds upon individual commitment, I started out from the very plausible assumption that people are often tempted to act in ways that do not support their long-term interests. Because of this limitation of our motivational integration – i.e. they illustrate that our currently predominant motivation does not always adequately reflect what is our long-term interests, there is an important functional role for commitment as a device which shields goal that are in our long-term interests from fluctuations in our short-term interests and passing impulses.

From this perspective, cases of individual and social commitment can be seen to overlap substantially with respect to the mechanisms which they engage to stabilize motivation to act towards valuable goals. They can also be seen as distinct but complementary sources of goal valuation, i.e. as providing distinct but complementary reasons why some goals are identified as be in in our long-term interests and accordingly worth shielding. To shed light on how these distinct but complementary sources of goal valuation relate to each other, we homed in on various stages along the

way from goal selection to goal completion. This enabled us to identify various individual and social factors which can come into play to increase goal valuation along the way through the progression – e.g. as a result of having selected, planned, initiated action towards the goal, etc.

As we saw, there are social reasons for valuing certain goals more than others, just as there are non-social reasons. But as we select particular goals, construct plans to achieve them, initiate action and persist towards goal completion, various social and non-social factors pile up and progressively buttress our commitment. Though our focus here has been on illuminating the ways in which social commitment builds upon individual commitment, we have also had occasion to observe many ways in which social commitment enhances individual commitment and creates new forms of individual commitment. It is surely no wonder that a species marked by its capacity for intricate long-term planning should also be a species in which cooperation is as pervasive as it is among humans.

References

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749-753.

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

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59-78.

Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing implicit commitment. *Review of Philosophy and Psychology*, *10*(4), 669-688.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology*, *66*.

Bratman, M. (1987). *Intention, plans, and practical reason* (Vol. 10). Cambridge, MA: Harvard University Press.

Chennells, M., & Michael, J. (2018). Effort and performance in a cooperative activity are boosted by perception of a partner's effort. *Scientific reports*, *8*(1), 1-9.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2), 193–201.

Dreisbach, G., & Haider, H. (2009). How task representations guide attention: further evidence for the shielding function of task sets. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 477.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical studies*, *144*(1), 167-187.

Heath, C. (1995). Escalation and de-escalation of commitment in response to sunk costs: The role of budgeting in mental accounting. *Organizational Behavior and Human Decision Processes*, *62*(1), 38-54.

Heintz, C., Celse, J., Giardini, F., & Max, S. (2015). Facing expectations : Those that we prefer to fulfil and those that we disregard. *Judgment and Decision Making*, 10(5), 442–455.

Heintz, C., Karabegovic, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in psychology*, *7*, 1503.

Hofmann, W., Friese, M., & Roefs, A. (2009). Three ways to resist temptation: The independent contributions of executive attention, inhibitory control, and affect regulation to the impulse control of eating behavior. *Journal of Experimental Social Psychology*, *45*(2), 431-435.

Hull, C. L. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, *39*(1), 25.

Kelly, T. (2004). Sunk costs, rationality, and acting for the sake of the past. *Noûs*, *38*(1), 60-85.

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research*, *43*(1), 39-58.

Löhr, G (In prep). Normative and non-normative uses of commitment.

Luce, R. D., & Raiffa, H. (1989). *Games and decisions: Introduction and critical survey*. Courier Corporation.

Michael, J., & Pacherie, E. (2015). On Commitments and Other Uncertainty Reduction Tools in Joint Action. *Journal of Social Ontology*, 1(1). https://doi.org/10.1515/jso-2014-0021

Michael, J., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, 6, 1968. https://doi.org/10.3389/fpsyg.2015.01968

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291.

Rachlin, H. (2016). Self-control based on soft commitment. *The Behavior Analyst*, *39*(2), 259-268.

Read, D. (2004). Intertemporal choice. *Blackwell handbook of judgment and decision making*, 424-443.

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901-908.

Robinson TE, Berridge KC. Addiction. *Annu Rev Psychol* 2003; 54:25 – 53.

Robinson TE, Berridge KC. The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Res Rev* 1993;18:247 – 91.

Rusch, H., & Luetge, C. (2016). Spillovers from coordination to cooperation: Evidence for the interdependence hypothesis? *Evolutionary Behavioral Sciences*, *10*(4), 284.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard university press.

Schrift, R. Y., & Parker, J. R. (2014). Staying the course: The option of doing nothing and its impact on postchoice persistence. *Psychological science*, *25*(3), 772-780.

Shah, J. Y., Friedman, R., & Kruglanski, A. W. (2002). Forgetting all else: on the antecedents and consequences of goal shielding. *Journal of personality and social psychology*, *83*(6), 1261.

Shpall, S. (2014). Moral and rational commitment. *Philosophy and Phenomenological Research*, *88*(1), 146-172.

Siegel, E., & Rachlin, H. (1995). Soft commitment: Self-control achieved by response persistence. *Journal of the experimental analysis of behavior*, *64*(2), 117-128.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*(5), 495-518.

Székely, M., & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner's effort. *Cognition*, 174, 37-42. ISO 690.

Theriault, J. E., Young, L., & Barrett, L. F. (2020). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*.

Walton, D. (2002). The sunk costs fallacy or argument from waste. *Argumentation*, *16*(4), 473-503.