10

Planning and Prefigurative Politics: The Nature of Freedom and the Possibility of Control

Bryce Huebner

Like most animals, humans will learn to associate actions with the rewards and punishments that typically accompany them. By mining past experience for information that will improve future predictions, we track fluctuations in the distribution and value of rewards, monitor changes in the probability of gains and losses, and produce moment-to-moment estimates of risk and uncertainty (Montague 2006). But while these backward-looking systems explain a surprising amount of human behavior, they make our ability to plan ahead, and exert forward-looking control over our behavior, somewhat mysterious.[1] Yet, before visiting a new city, I often search for information about the best cafes, and make plans to visit some. By focusing my attention on some cafes, and removing others from further consideration, I probably miss some hidden gems. But planning ahead also reduces the uncertainty inherent in finding coffee in an unfamiliar city, and minimizes the risk of drinking bad coffee. When I travel with friends, we often plan together. This prevents forms of indecision and dispute that would prevent us from drinking coffee; and sometimes it allows us to uncover sources of excellent coffee that I would have missed.

Coffee is important. But such phenomena are the tip of a much larger iceberg. Many humans can imagine novel possibilities, decide which goals to pursue, and decide which means to employ in pursing them; and as a result, we can improve our situation, instead of just acting to minimize the effects of our contingent learning histories (Dennett 1995, 377ff; 2003, 266-68). The ability to look ahead gives us control over our actions, and it does so by opening up elbowroom to act in accordance with our values and ideals.[2] My aim in this chapter is to clarify some of Daniel Dennett's significant insights about these distinctively human forms of freedom. Specifically, I want to explore his claim that moral agency often depends on a little help from our friends (Dennett 2003, 135). I argue that by deliberating together about collective actions we can open up forms of prefigurative practice that change the available possibilities for action. But I build up to this claim slowly, and perhaps surprisingly, by 'turning the knobs' on Michael Bratman's (2014) theory of planning agency.

1. Intentions as planning states

Bratman (2014, 27) conceives of intentions as planning states, which play a critical role in our "internally organized temporally extended agency and to our associated abilities to achieve complex goals across time, especially given our cognitive limitations". Intentions terminate reasoning about which ends to pursue, prompt reasoning about the best means of pursuing them,

[1] In this chapter, I focus primarily on the possibility that explicit, propositionally articulated thought and speech can play a role in action guidance. I think that such forms of action guidance are rare, and difficult to sustain in the face of ongoing subpersonal processing and unconsciously computed expectations. But I do think that such states can facilitate forms of cognitive control, and I hope to offer some insight into how they are able to do so.

[2] I would like to thank the members of a class on "Free Will, Intention, and Responsibility" at the Jessup Correctional Institute in Maryland for helping me to see the power of this idea.

and serve a coordinative role in temporally extended and collaborative actions. And as Bratman notes, intention-governed cognition typically unfolds in two phases: we first form a partially abstract plan; then we anchor it to a particular situation, specifying what must be done to achieve our ends, given our current circumstances. This makes our plans flexible, yet binding, allowing them to structure our decisions and guide our ongoing behavior.

In forming intentions, we should be sensitive to information that might affect the feasibility and desirability of the actions we are considering. But this requires attending to norms of consistency and coherence: we aim to make our plans internally consistent; we aim to make them consistent with our beliefs about the world; and we aim to make them consistent with our other plans and projects. In forming an intention to teach a course on the evolution of language next Spring, for example, I commit to avoiding conflicts with other courses and to updating my plan if I learn of new theoretical or scientific perspectives; this ensures that my plan remains consistent with available information. I also commit to revising or abandoning my plan if I learn that it is internally inconsistent (e.g., because the issues require too much background knowledge for the students), unlikely to succeed (e.g., because students are uninterested), or in conflict with my broader network of plans and projects (e.g., if I will be unable to meet important writing deadlines if I try to develop a new class). Finally, forming this intention should lead me to stop thinking about what to teach, and to start thinking about the best means of teaching. And by settling on plausible means for executing my plan, I provide myself with the time to do the necessary reading, and the time to construct a coherent course narrative.

In general, forward-looking plans evoke present-directed intentions, which motivate us to pursue our ends, and to continue to do so until we reach them. They shift our attention to the guidance of behavior, in ways that allow us to maximize efficiency and effectiveness (Pacherie 2013, 150). And in this respect, they impose a cognitive filter on thought and behavior, highlighting action-relevant information and action-relevant facts about our current situation; and as a result, we pursue our intended actions unless we learn that they will have problematic collateral effects. Having formed an intention to drink some coffee when I finish writing this paragraph, I will typically take action to do so. But I will abort my plan if I realize that the only coffee around is Ruth's soy cortado; while my current informational state is goal directed, I remain sensitive to inconsistencies with other goals (e.g., not making my partner angry). Likewise, I will abandon my plan to teach particular articles if I learn that students don't have the background to understand pushdown architectures. But in general, unless we run into trouble, our plans guide behavior in accordance with our forward-looking plans because they represent a situation we want to bring about, as well as a path to achieving our goal that we have committed to (Bratman 2008, 53).

A wide range of psychological data confirm that we think about things differently when we deliberate about which ends to pursue, and when we are try to find the best means of implementing our plans and organizing our ongoing behavior (see Gollwitzer 2012 for a review). Deciding which ends to pursue tends to evoke a *deliberative mindset*, in which attention becomes focused on information about the feasibility and desirability of different options. In this mindset, we remain more open to alternative possibilities, more realistic about the prospects of success, and more sensitive to the pros and cons of different options. People in deliberative mindsets also tend to be more accurate in their assessments of how much control they have over future actions and outcomes, including the risk of being in a car accident, the risk of becoming depressed, the risk of developing a drinking problem, and the risk of being mugged (Taylor & Gollwitzer 1995). In part, this is because deliberative mindsets evoke heightened receptivity to potentially relevant information, leading us to look for different causal explanations of the things that happen in the world (Gollwitzer 2012, 537). This makes sense. A person who is trying to decide what to do cannot know in advance where their decision will take them; so being receptive to information is an appropriate

and functional task solution (Gollwitzer 2012, 528). By being careful not to dismiss information that might be useful later, people in a deliberative mindset can thus become more sensitive to factors that affect the coherence of their ends, and somewhat more accurate in their estimations of how likely they are to reach their goals.

Once people settle on a goal, they tend to focus on goal-relevant information while ignoring or downplaying goal-irrelevant information (Gollwitzer 2012, 534). In an *implementation mindset*, it makes sense to focus on the best means of achieve our goals; but this leads people to be more optimistic about their chances of success, and more partial in their judgments about the desirability of the ends they pursue. People in such mindsets tend to over-estimate the control they have over various outcomes, and underestimate their vulnerability to various risks (Taylor & Gollwitzer 1995). By focusing on information related to the achievement of their ends, and ignoring other sources of information, we can more rapidly determine the best way to achieve our goals, and shield ourselves from the distractions imposed by competing considerations (Gollwitzer 2012, 528). As a result, we become more resolute in the pursuit of our goals—especially where the perceived feasibility of a task is low, but the desirability is high, or vice versa (Brandstätter & Frank 2002). But this comes at a cost: we can easily miss important information if it is not within the range of information to which we are currently attentive.

2. Foreknowledge and freedom

Dennett (2003, 251) often argues that one of the most distinctive facts about human agency is that we can make requests of ourselves, and at least sometimes we comply with them. This requires more than just acting in accordance with a command. My cat might jump off the table when I tell her to get away from my coffee, but she doesn't understand the sounds I produce. And when she ignores those sounds, she is not declining my requests. Furthermore, she cannot make requests of herself; she cannot decide to limit her food consumption to improve her jumps; and more generally, she cannot make explicit plans and use them to guide behavior. By contrast, when we intend to do something, our plans place normative constraints on behavior, and deviations from our plans call for reasoned explanation (Bratman 2008). Like failures to comply with accepted requests, the failure to follow through on our commitments feels bad, and calls for a justification.

Of course, failures to follow through on our plans can arise in many different ways, and each calls for a different response. Follow through is difficult when we are tired, busy, or overwhelmed; and even minor stress can lead us to abandon computationally taxing forms of explicit planning in favor of habitual and Pavlovian forms of behavior (Crockett 2013; Schwabe & Wolf 2013). But in such cases, our failures are mechanical, and a plausible explanation can proceed from within the design stance. We can also change our minds before we act, even when we have settled on a plan. My plan to make a cup of coffee, for example, may fail to trigger a present-directed intention when I decide that I don't have time to do so, or when I decide that a 6th cup isn't necessary. But there are many other reasons why we don't follow through on our plans, including picking tasks that were too intellectually or physically difficult to complete, or too conceptually or physically distant from our current situation. These are paradigmatically rational decisions, and I can explain and justify them, both to myself and to others. But sometimes I just lack the "willpower to get started, to stay on track, to select instrumental means, and to act effectively" (Gollwitzer 2014, 305). These failings are agential, but not rationally justifiable, and our ability to control them reveals an important sense in which foreknowledge can yield control over our ongoing behavior.

By looking ahead, and imposing forward-looking constraints on behavior, we can sometimes find better ways to act in accordance with our goals and values. Precommitting to an action can decrease the likelihood of pursuing immediate rewards at a cost to our future interests (Ainslie

2001), and it can decrease the likelihood of cheating or shirking when the going gets tough (Schelling 1966). As Dennett (2003, 207) notes, precommitment raises the stakes, and "changes the task of self-control we confront". Precommitting to an action shields us against temptations that might lead us to abandon our chosen ends (Crockett et al 2013). And since intentions are conduct-controlling pro-attitudes that we are inclined to retain without reconsideration, they should be able help us solve more complex intrapersonal commitment problems (Bratman 1987, 20). And they do.[3]

Precommitting to a plan delegates control over the initiation of an action to a situational cue, creating a strong associative representation that links that cue to a relevant response (Gollwitzer 2012, 537). By specifying a precise, situation-specific action-plan, we automate the translation of future-directed intentions into present-directed intentions (Adriaanse et all 2011b; Crockett et al 2013; Gollwitzer 1999). The method for doing this is surprisingly simple. Rehearsing 'implementation intentions', simple if-then plan that specify precise trigger-cues and behavioral responses, instigates a form of reflexive action control that shields behavior from temptation and distraction, and counteracts the effects of habitual behavior patterns (Gilbert et al 2009). I can increase the likelihood of taking my multivitamin each morning, for example, by forming the following plan: "When I have my morning coffee, I will take my multivitamin". While I might lack the processing power to remember to take my multivitamin in the morning, this intention creates a significant association between a particular situation and a particular action, making it easier to remember to act, by increasing the salience of my intention in the relevant context (Gollwitzer 2014). Similar intentions can affect the ways that I think, my affective responses, and the behavioral dispositions I adopt; they can also reduce disruptive influences that originate from innate action tendencies, learned habitual responses, behavior priming, or entrenched social prejudice (Gollwitzer 2014, 311). By imagining a future where we act in a particular way, we can recruit associative mechanisms to guide our behavior in accordance with the plans we form.

Similar effects arise in many cases, and a large meta-analysis has revealed substantial effects of implementation intentions on goal-attainment across domains as diverse as consumer decisions, academic achievement, health-relevant behavior, and morally significant phenomena like being more egalitarian and more pro-social (d=.61; Gollwitzer & Sheeran 2006); and more targeted meta-analyses have confirmed these effects for behavior relevant to healthy eating and physical activity (Adriaanse et al 2011a; Belanger-Gravel 2013). Most surprisingly, implementation intentions can moderate the perceptual tendency to mistake a cellphone for a weapon in a racialized context (Mendoza et al 2010; Webb et al 2012). By rehearsing the plan "If I see a Black person, I will think 'safe'" people can bring their reflexive behavior into line with antecedently held egalitarian goals. But as powerful as they are, implementation intentions only guide behavior where we already have strong commitments to particular goals. Since they operate by linking high-level plans to target cues, they can only yield present-directed intentions were we are disposed to prefer the actions we are attempting to control. Put differently, implementation intentions allow us to maintain top-down control over behavior by using foreknowledge of future contingencies to arrange associative thought in ways that prevent us from pursuing worse options though we know better ones are available (cf., Spinoza 1677/2002). Consequently, the strongest effects of implementation intentions arise when people use cues they already consider critical to an action, and link these cues to a behavioral response they already believe to be relevant to the achievement of that goal (Gollwitzer 2012, 541).[4]

[3] As John Gavazzi (p.c.) reminds me, precommitment strategies also play a prominent role in clinical environments, where people are trying to change specific behaviors such as overeating or smoking.

[4] While implementation intentions only succeed where a goal is perceived as important, and preferable to the existing alternatives, there are ways of affecting these states as well. By imagining a desired future, and reflecting on the facts that

3. Planning and cognitive architecture

Neuroscientists have long known that even highly routinized tasks, like picking up a coffee mug, require integrating goal-based representations, perceptual inputs, and ongoing proprioceptive and evaluative feedback, to yield representations encoded for use by motor systems, control loops, and emulator circuits (Akins 1996, 354; Mahon & Caramazza 2008). As Elisabeth Pacherie (2006; 2013) argues, this fact suggests an important role for motor-intentions in the guidance of behavior. Motor mechanisms don't require the guidance of explicit plans; they operate reflexively by minimizing discrepancies between predicted and actual motor states, and making micro-adjustments whenever such discrepancies arise. But just as importantly, these systems are embedded in a computational hierarchy, and integrated with systems that constrain their behavior in some cases. At the top of this hierarchy, forward-looking intentions, structured as explicit plans, lead us to pursue goals that we have settled upon rationally; at the mid-level, present-directed intentions arise as we flesh-out situation-relevant specifications of those plans—guiding behavior in accordance with our forward-directed intentions; and at the bottom, motor-intentions initiate and guide behavior in real-time, facilitating moment-to-moment control over our ongoing action.

There is a rapidly expanding consensus that the human brain is a hierarchically organized predictive machine, which consists of numerous error-driven learning systems that each have their own tasks, and their own models of those tasks (Dennett 2015; Friston 2009; Howhy 2013; Rescorla 1988). Systems closer to the bottom of this hierarchy traffic in more precise sources of information, and processing becomes more abstract and more conceptual toward the top of the hierarchy. But although each system provides input to the system above it, the goal of cognitive processing is not to build a more useful representation of the world as information is propagated upward. Instead, each system attempts to predict the inputs it will receive, given its model of the world. These predictions flow downward, resolving ambiguous data without additional search. At the same time, each system generates error-signals when surprising data are encountered; these error-signals flow upward, recruiting new top-down predictions to better accommodate the incoming data. Over time, this bi-directional flow of information allows the brain to search for "the linked set of hypotheses that, given the current raw sensory data, make that data most probable" (Clark this volume).

Andy Clark has shown that the predictive coding framework can clarify the extent to which top-down predictions affect perceptual experiences. As we move about the world, we extract and complete perceptual patterns in accordance with our perceptual expectations. And often, these expectations allow us to act even though the incoming perceptual signal is noisy, and even though there isn't enough incoming information to guide situation-relevant action. Lisa Feldman Barrett (2014) has argued that top-down expectations also have a significant impact on the emotions we experience, as well as the action-tendencies that are recruited as we move though the world. And I've argued elsewhere that top-down expectations often impact our ability to conceptualize possibilities for socially significant action (Huebner 2016). And here, I focus on expectations that take the form of planning states.

A highly salient example of this occurs when expectations about violent Black males and the likelihood of violent crimes in particular neighborhoods guide action-oriented perception. A person passing through a neighborhood that evokes these expectations will experience increased awareness

---

stand in the way of reaching that future, people can enhance the cognitive relevance of the desirable features of an action, and in this way, they can highlight the feasibility of their plan (Oettingen & Gollwitzer 2010). I return to this point in the final section.

of potential threats, negative affective valence, and enhanced accessibility of stored associations between race and violence; this will often trigger the construction of action-plans designed to prepare this person to navigate potential threats (Wilson-Mendenhall et al 2011), and make it more likely that they will *see* a cell-phone as a gun (Barrett 2015). In this situation, ambiguous visual stimuli (e.g., a barely glimpsed cell phone) are more likely to be resolved in accordance with racialized expectations, and thus will feedback into action-oriented processing, thereby increasing the likelihood of acting as though the innocuous object is a threat. This claim might seem surprising if we assume that incoming data fully determine what we see and what we are motivated to do. But conceptualizing the brain as a behavioral guidance system, which is designed to use ambiguous data to navigate a dangerous and dynamic world, predicts that higher-level expectations will often be used to resolve perceptual ambiguities in favor of expectations about what is likely to happen next; where we expect a pattern, the brain will try to complete it. And sometimes the results are awful!

But what role do planning states play in this kind of processing? As I noted above, implementation intentions can have an impact on everything from motivation, to ways of thinking and affective responses, in ways that yield robust behavioral dispositions—and they can do so even where perceptual cues are seen briefly or tracked subconsciously (Gollwitzer 2014, 308). In forming such intentions, we generate new top-level expectations, which explicitly link a target cue with a relevant response. These expectations have conceptual structure, but they also provide a top-down signal that can guide the response of lower-level systems to ambiguous sensory inputs. Implementation intentions serve as models, which can be fed downward to lower-level systems (Huebner 2016; Kim et al 2011); and so long as the error-signals that lower-level systems compute do not routinely violate these expectations, these models, which are constructed at the top level of the tower of generate-and-test, will continue to constitute part of the overall hypothesis that best approximates the incoming data. Over time, unless error-signals require adopting alternative hypotheses, cognition should converge on a linked set of hypotheses that make actions cued by implementation intentions most plausible. But as George Ainslie (2001) suggests, failures to act in accordance with our plans should be disastrous, as they will cause this linked set of hypotheses to collapse; and where the world routinely provides perceptual information that contradicts our expectations, they should shift to become consistent with the world, even if they are less consistent with our top-level goals. So our intentions sit atop the tower of generate and test, serving as top-level expectations that can guide behavior where doing so is necessary.

In light of these claims, I maintain that Dennett is right about several things: multiple processes, operating at multiple levels, guide ongoing behavior; planning depends on explicit representations that lie at the peak of the tower of generate-and-test; explicit representations only need to arise at the top-level of the tower of generate-and-test; and plans allows us to enact forms of forward-looking control by using socially-situated conceptual knowledge to re-shape our responses to the world from the top-down. By binding our understanding of the world to particular situations and activities, we can rely on perceptual, evaluative, and motor representations to guide our moment-to-moment behavior; but we also rely on "revisable commitments to policies and attitudes that will shape responses that must be delivered too swiftly to be reflectively considered in the heat of action" (Dennett 2003, 238). I thus contend that we should accept something like the following view of human agency (though the details may shift as we learn more about the hierarchical structures that are operative in the guidance of goal-directed behavior):

- Future-directed intentions, structured as planning states, yield high-level expectations and allow us to impose goal-directed structure on our actions. Our future-directed intentions

typically evoke present-directed intentions because they provide top-down pressure on the computational cascade that triggers goal-directed behavior.

- Absent-minded and *akratic* behavior arise where we default to habitual or innate responses patterns, and where encounters with the world lead us to pursue worse options though we know better ones available (Spinoza 1677/2002). But with foreknowledge of the conditions under which these failures of agency are likely to occur, we can use top-level intentions to prevent such failures (using implementation intentions and precommitment strategies).
- Present-directed intentions can also arise at the mid-level of computational processing, even without future-directed intentions to guide them. This will produce spontaneous intentional behavior, or rational forms of habitual behavior (Bratman 1987, 119ff; Tollefsen 2014). Sometimes free-floating rationales, which are not represented explicitly, guide such actions; and sometimes internalized rational expectations that allow us to navigate our social world guide such actions habitually (as we'll see below, this yield problems for us as agents).
- Finally, motor-intentions can sometimes be produced without evoking present-directed intentions, and this will yield forms of goal-consistent behavior without rational control. Such behavior should be experienced as reflexive and automatic, but post-hoc rationalization may nonetheless allow us to treat such behavior as resulting from existing plans, thereby yielding illusions of conscious control (Pacherie 2006).

Unfortunately, I worry that the kind of freedom that foreknowledge provides on this picture isn't quite as robust as we might hope. And I am less sanguine than Dennett (2003, 239) about the claim that we are the "authors and executors of these policies, even though they are compiled from parts we can only indirectly monitor and control".

4. A Killjoy Interlude

While implementation intentions provide evidence of distinctively human forms of action-guidance (Holton 2009), the reason for their success also reveals their most troubling limitation. Expectations can significantly affect perception and action, but they aren't created out of whole cloth. We acquire particular expectations though our encounters with the world; and evaluate and select actions on the basis of expected rewards, and evaluative facts about our current state and current motivations (Dehaene & Changeux 2000; Montague 2006; Polania et al 2015). But we live in a world that is thick with structural racism, sexism, ableism, trans*phobia, and xenophobia. So as we watch TV and films, read novels and blogs, and walk through familiar and unfamiliar spaces, we are bombarded with statistical 'evidence' that fosters the construction of exclusionary assumptions; far more troublingly, it's rare for most of us to encounter situations that would recruit and sustain the anti-kyriarchical expectations that many of us hope to formulate.

These social structures impact our highest-level values and ideals, both the ones that guide our behavior unconsciously, and the ones that we consciously avow. We see the latter when philosophers express a preference for particular questions and methodologies, treat certain things as signal and others as noise, and assume that answers must take a particular form. As Dennett has long argued, people often acquire these attitudes as part of their training as philosophers, leading to patterns of convergence in our assumptions about which questions deserve answers, and which answers are viable. But the problem runs much deeper. As Nathaniel ~~Coleman~~ (2015) argues, the discipline of philosophy has been *whitewashed*, and contemporary Anglo-European philosophy displays what John Dewey (1930, 26) referred to as the deplorable deadness of imagination inherent in supposing that "that philosophy will indefinitely revolve within the scope of the problems and

systems that two thousand years of European history have bequeathed to us". The institutional structures that govern academia foster expectations grounded on racialized assumptions that are rarely acknowledged, and less often challenged. Similar problems arise in most areas of our cognitive and social lives (cf., Dennett in prep). Because we attune to social practices, our high-level attitudes and low-level reactions entrain to statistically prevalent norms and practices. This leads to stable institutional structures against which future attitudes and behavior can attune; and as a result, judgments about which plans are feasible and desirable tend to be structured around cognitive biases that have become entrenched in social norms and practices that guide behavior in ways that are beyond cognitive control.

Even worse, attempts to transcend our biases, by constructing expectations that run contrary to dominant social norms, typically give way to actions that accord with the norms and practices we are trying to overcome. This happens as actions driven by counter-social expectations are met with feedback suggesting that we are making mistakes (Klucharev et al. 2009, 2011); and where error-signals continually arise, new hypotheses are recruited to make our behavior more consistent with the world we inhabit. Put much too simply, as the brain searches for the linked set of hypotheses that make incoming data most plausible, our expectations will shift toward statistically common patterns that we hope to overcome. Our degrees of freedom are therefore socially limited, and the freedom we have to plan ahead is constrained by the norms that govern our social lives. Put much too bluntly, we are free to conform in the long run, even though we can resist in the short run.

As far as I can tell, this is an implication of Dennett's (2003, 273) hypothesis that a "proper human self is the largely unwitting creation of an interpersonal design process in which we encourage small children to become communicators and, in particular, to join our practice of asking for and giving reasons, and then reasoning about what to do and why". This is not to deny his claim that we are "capable of altering course at any point, abandoning goals, switching allegiances, forming cabals and then betraying them, and so forth" (Dennett 2003, 155). But we can only do so on the basis of goals and values that we have acquired through our interactions with the world. And where problematic patterns are pervasive, and where we act in ways that feel right because they are statistically regular, we will only hit upon normatively valuable practices by accident or luck.[5]

Nonetheless, there is something right about Dennett's claim that we have more degrees of freedom than other animals do. We can get stuck on the local peaks of an adaptive landscape, and we rarely consider the possibility of better options. But we do have the capacity to imagine options that aren't currently available (Dennett 2003, 267). As I see it, the problem we face is a Darwinian problem: it is hard to sustain novelty. I have argued elsewhere that doing so requires building a world around preferable values and ideals (Huebner 2016). But if the arguments I have advanced here are roughly right, judgments about what kind of world is preferable will also be governed by the social world we inhabit. Nonetheless, I believe that *planning together* can sometimes open up new possibilities. By working together we can imagine another world, and attune to local ways of thinking that can prevent the kinds of cognitive backsliding to which Bayesian thinkers often fall prey. My aim in the remainder of this chapter is to explain how this is possible.

5. Using cognitive prosthetics

Let's start with a banal case of plan-driven behavior: making an excellent cup of coffee with a v60. This plan is complex, and to successfully execute it I must: place a v60 atop a mug; place a filter inside the mug; boil water; rinse the filter; discard the rinse water; grind the proper quantity of

[5] I take this to be one of Plato's insights in Gorgias, and the core of Spinoza's worries of regarding human bondage.

beans and place them in the filter; make a divot in the center of the grounds; pour an ounce of near-boiling water over the grounds and let them bloom for 30 seconds; and pour water over the grounds, in concentric circles, until the mug is filled with delicious coffee. Some of these steps are more optional than others. I can decide not to rinse the filter, make a divot, or let the grounds bloom (though these choices will affect the quality of my coffee). And I can skip these steps without losing track of which steps must still be executed, even where they have always been carried out in the past (Fitch & Martins 2014). Furthermore, I can boil the water or grind the beans first, with little impact on the remaining steps. These steps are required, but their order is somewhat optional; they must only occur prior to the bloom. But I cannot pour the water unless I have placed the filter and beans in the v60; like many actions, this one is hierarchically structured, and some actions must be carried out before others. Finally, I can pause after completing any step prior to the bloom, and resume without compromising success. These facts seem to suggest that I am able to mark where I am in the process, and track the dependencies between various sub-tasks; and this seems to require complex internal capacities for storing and manipulating internal representations (perhaps using something like a pushdown stack architecture; cf., Fitch & Martins 2014, 96).[6]

But before we get too comfortable with claims about such forms of internal guidance, we must note that we often supplement our representational capacities with cognitive prosthetics; and our plans become sparser as the information in our environment becomes more richly structured (Simon 1969). As Zoe Drayson and Andy Clark (forthcoming) argue, people with Alzheimer's often rely on this fact to compensate for internal memory deficits. They enrich their material and social environment by placing notes around their houses specifying what to do and when; they label things; they set up reminders to organize their behavior; and they rely on loved ones to help navigate their lives. This social and material scaffolding also helps them to make plans, and to find ways of living more autonomously. But this is also the standard situation for neurotypical people; and forms of social and material scaffolding play a significant role in our ability to remember the past, plan for the future, and consider alternative possibilities (Kosslyn 2006).[7]

Consider Barika's use a list of processes, written on a sheet of paper, to guide her coffee-making behavior. At each step, she must check the sheet, and carry out the next step. This makes her action rigid and inflexible, but the structure of her informational environment minimizes the information she must remember. Depending on her working memory capacities, and her ability to think clearly in morning, this may be the best way to successfully make coffee. She still needs to

[6] Intriguingly, sequential planning can also be compromised while the ability to carry out habitual actions, and even component actions is preserved. In rare cases, *action disorganization syndrome* can lead a person to omit tasks, do them in the wrong order, or perseverate on a single task (Humphreys et al 2001; Jackendoff 2007). They might forget to put the filter in the v60, pour water directly into the mug, or continue to pour water even when the cup is full. But while forward-looking plans do seem to be realized by systems beyond those employed required for habitual learning (Lashley 1951), there is an ongoing debate over whether planning requires systems that construct hierarchical representations (Fitch & Martins 2014; Rosenbaum et al 2007), or whether Bayesian systems dedicated to probabilistic inference suffice (Botvinick & Toussaint 2012). At this point, the underlying architecture remains unclear. Thanks are due to Joey Jebari for pushing me to clarify this issue.

[7] An initial attempt at understanding the kinds of neural processing that make cognitive offloading possible has recently been carried out by Julia Landsiedel and Sam Gilbert (2015). They found increased BOLD response in a network of 'task-positive' regions when people remembered delayed intentions, as well decreased response in the so-called default network. But when participants set external reminders for themselves, the deactivation of the default network was strongly reduced where there was a larger memory load, even though there was no parallel reduction in task-positive activation. They suggest that this is because medial rostral PFC activity is associated with externally cued rather than self-initiated activity. For our purposes, what is most interesting in these data is the suggestion that internal and external sources of information are being integrated to guide online behavior.

track where she is in the task, and she still needs to flesh out details of her plan to suit her current situation. But by planning to have this list of processes on-hand, Barika can increase her degrees of freedom, by changing the world instead of changing herself. Of course, changes to the self will often happen down-stream. When I acquired a rakweh in Lebanon, I had to search the Internet for instructions, and follow them rigidly until I could make coffee with this simple device. Over the course of many weeks, I began to tweak the plan I had read about, and found ways to manipulate the component actions to develop a more flexible skill for making excellent Lebanese coffee (both with and without cardamom). But the initial planning phase depended on externally stored representations. And in many cases the representations we exploit remain outside the boundaries of skin and skull. Many of us now exploit the Internet as a cognitive prosthetic for developing forward-looking plans, and guiding online behavior; as a result, we tend to think about computers when we are presented with hard questions, and we are less likely to encode information if we believe it will be available on-line (Sparrow et al 2011). This too can open up degrees of freedom that would otherwise be unavailable.

More interestingly, we can rely on other people to scaffold our ongoing behavior; and collaboration can also expand our degrees of freedom. Suppose two people are trying to navigate an unfamiliar city, in the dark, using a small map they have received at a conference (Gross 2012). Zoe struggles with spatial reasoning, and she can't quite figure out the relationship between the map and the streets; Phred has a hard time reading, and she has difficulties making out the words that are written on the map. But Zoe and Phred also have complementary capacities. Zoe can read the map as well as the street signs, and Phred can understand the spatial relations on the map, and their relationship to the streets they are walking along. By talking to one another, and developing meshing sub-plans to guide cooperative behavior, they can successfully navigate the city *together*.[8]

Bratman (2014, 32) argues that joint-intentions tend to arise in collaborative groups in which people (1) intend to act together, (2) on the basis of beliefs about what the others intend, (3) where their meshing sub-plans are common knowledge, and (4) where they are mutually responsive to one another's plans and actions. In this case, Zoe and Phred might both intend to get back to their hotel, with the other, in a way that's consistent with their distinctive plans and abilities. Each of them might believe that they will get to the hotel, as long as the other follows through on their plans; and as they walk, their joint intention help them respond to what the other says and does, and prepare them to update their plans if anything goes awry. If so, their joint-intention will allow them to achieve ends that neither could achieve easily on her own. Bratman (2014, 42) also argues that intending to act together triggers an inferential chain, which leads us to form intentions to play particular roles in bringing about joint-activities. For example, if Zoe and Phred make a plan to get coffee at 15:00, Zoe will tend to form the intention to drink coffee with Phred at 15:00, and this will trigger actions directed toward bringing it about that they drink coffee together. As they plan, she should be sensitive to norms of *interpersonal consistency*, which will focus her attention on considerations that affect the feasibility and desirability of their shared end (Bratman 2014, 29). And as they act, Zoe's focus should shift to the role she plays in bringing about their joint-activity (Bratman 2014, 64). Where things go well, this should also lead Zoe to provide assistance to Phred where doing so is necessary; and it should lead her to make adjustments to her plan in light of

[8] The claim that agency and autonomy depend on or relationships with others, institutional structures of social power, and the forms of self-understanding that they engender, is far from novel (Christman 2004; Holroyd 2011; Kukla 2005; MacKenzie 2014; MacKenzie & Stoljar 2000; McLeod 2002). But these issues have yet to take hold in discussions of the cognitive science of agency. In future work, I hope to flesh out the points of contact between accounts of relational autonomy and the cognitive science of agency—for now, this must remain a promissory note.

Phred's behavior. Bratman (2014, 89) argues that these forms of social responsiveness result from facts about Zoe's plans, given their social content, and given the demands of interpersonal consistency these plans entail. And this is consistent with the model of precommitment and planning that I discussed above; intending that *we act together* generates a high-level expectation, which can guide individual behavior in ways that constitute collective action.

6. Acting together

We must proceed carefully in thinking about the value of acting together. After all, statistical learning mechanisms often suffice to calibrate behavior against the structure of our environment, and to prioritize competing sources of sensorimotor information (Dehaene & Changeaux 2010; Shea et al 2014). This is even true where multiple agents are making decisions in parallel. Imagine a busy cafe with overlapping workspaces. Without compromising the flow of high-quality coffee, Mathew might need to put a cup in a space that Terry was keeping clear, and Rachel might need to get some beans from a drawer behind Mathew. These baristas must continually update their motor-intentions and adjust their present-directed intentions to accord with their interactive context. And to do so, each must track their own actions, the actions of other baristas, and the unfolding of their joint-actions; and each of them must dynamically update their behavior and intentions in light of changes in this situation (Tollefsen 2014, 15). But if baristas had to track all of this information explicitly, real-time coordination would be impossible (Kirsh 2006); and this point generalizes, the flow of information in many real-world environments is often rapid and sparse, making explicit planning impossible (Blomberg in press; Tollefsen 2005).

Fortunately, Deborah Tollefsen & Rick Dale (2012) have shown how real-time coordination can be sustained by subpersonal mechanisms that track multiple sources of information in parallel to bring the behavior of interacting agents into alignment. We often treat our social environment as a resource to be probed with epistemically directed actions, and adjust our motor-states against the stream of data evoked by these probes (Kirsh & Maglio 1994; Friston 2009, 295). In interactive contexts, the behavior of others partially constitutes this environment, and we can probe others, track forms of behavioral feedback (including facial expressions, body postures, gestures, and more), and adjust our own behavior reflexively and automatically (Tollefsen & Dale 2012, 392). Interpersonal epistemic action thus allows us to formulate present-directed intentions that allow us to act together, fluidly and dynamically, without explicit plans or shared representations (Tollefsen 2014, 21; Wilutzky 2015). Where individual differences in the construal of a problem or a situation are irrelevant, these present-directed intentions can arise as mid-level representations, guided by free-floating rationales that are not explicitly represented. This is probably the normal context in which human behavior unfolds, as expectations are rapidly and continuously updated to ready us for situation-relevant forms of thought and action (Barrett 2014). But, social interactions recruit social expectations; and sometimes, as we conceptualize ourselves as participants in joint-activities, forms of joint-attention and explicit planning are recruited to navigate social situations.

As our behavior aligns with the behavior of social partners, we often exhibit more interactive forms of joint-agency, and this triggers "more expressions of working together, feelings of solidarity, and so on" (Tollefsen & Dale 2012, 402). Over time, this can bring our explicit attitudes into alignment, and make us more sensitive to the options available to us as a group; as a result, we may be led to think of things that are less likely to occur to observers of a collaborative activity. Acting together can focus attention on the intentions, reasons, and emotions of others, allowing us to rely on different sources of information in the construction of explicit plans (Gallotti & Frith 2013, 162). To reason and plan together, people must "survey and convey to others their own thoughts, feelings, memories, and imaginings. This requires not only representational capacities, for which

consciousness is surely not needed, but meta-representations that we can make publicly available—including embedded if–then meta-representations of what one would think, feel, or seek to do under hypothetical circumstances" (Seligman et al 2013, 130). And as Dennett argues, the use of explicitly represented mental states can open up novel possibilities for action control and action guidance. We are now in a position to see how this occurs (cf., Sie 2014 for a set of broadly similar suggestions).

The production of explicit representations satisfies a supra-personal cognitive demand: explicitly represented mental states can be broadcast to others, and this allows us to bring action-guiding systems into alignment across differences in learning histories, and differences in the construal of the current situation; by broadcasting explicit representations, we can triangulate cooperative behavior across notional worlds (Shea et al 2014). As Bayesian learners, we constantly attempt to generate the best model of the world we inhabit; in the process, we attune to the statistical contingencies we have encountered. Each interaction with the world shapes our experiences and expectations. And since we encounter subtly different aspects of the world, in different affective states, with different expectations, the models we construct are likely to diverge in multiple ways that will go unnoticed without careful heterophenomenological investigations. The problem is compounded by the fact that biological differences in impulsivity, risk aversion, reward sensitivity, and perceptual acuity will have an impact on the features of the world that we each attend to, yielding complex differences in our understandings of what is possible and what is impossible in the world that we all inhabit. So we must 'talk' to one another to coordinate across differences in notional worlds, as well as differences in moods, expectations, and biases; but our success in doing so will often vary as a result of our openness to others, and our willingness to reconsider our initial construal of a situation.

I don't mean to oversell this point, as many differences between our notional worlds are irrelevant to individual and joint activities. Nonhuman animals can act together without relying on explicit representations (Couzin 2009), and so can children who do not possess robust capacities for mutual understanding (Tollefsen 2005); strangers can push cars out of the snow together without planning to do so; and protests can emerge spontaneously, among people who only have weak expectations about what others will do (Kutz 2000). Differences between notional worlds are also minor enough in many cases that they have no noticeable effect on attempts to navigate a shared world. We all live in the same world, and we attune to many similar reward contingencies and statistical regularities. This can create notional worlds that replicate hegemonic ideologies, and that shape evaluative judgments to accord with dominant power structures (Huebner 2016).[9] But occasionally, differences in notional worlds do make a difference, and explicit representations must be recruited to understand someone who experiences the world differently from us. This happens when we interact with someone with different cultural assumptions, a different political ideology, or different patterns of engagement with a lived environment. And it happens where we want to collaborate, in a complex decision-space where we are unsure whether our collaborator views the current situation in the same way we do (Shea et al 2014, 188). By making representations of our states available to others (using gestural or verbal reports) we can create publicly accessible signals that can be evaluated for significance relative to joint-activities from multiple perspectives.

[9] Philosophers should pay more attention to the complex interactions between the material structure of our world, the structure of notional worlds, and the capacity to act freely. These discussions have been at the center of research on Disability; they were essential to Frantz Fanon's discussion of colonial power and colonized psychologies; and they were the foundation of W.E.B. DuBois's discussion of double-consciousness. A recent paper by Oulfemi Taiwo (in prep) helps to clarify how these factors affect rational control and agency; and I hope to pursue these issues in my next book.

Shared representations can impose structure on the world that wasn't there previously, by 'freezing' contents to serve as conceptual anchors in a sea of dynamic thought (Clark 1996). And once they are present in collaborative contexts, explicit representations can be used to synchronically to coordinate ongoing group behavior, or diachronically to facilitate the revision and adjustment of previously stored or expressed representations (Shea et al 2014). As a result of these forms of supra-personal cognitive control, groups whose members communicate are routinely able to outperform groups whose members do not, in cases as diverse as mock-jury deliberations, perceptual and motor tasks, and tennis games (Shea et al 2014, 189). More intriguingly, these forms of mutual responsiveness can help groups realize things that individual members would have missed, and they can increase the likelihood of revising or abandoning failed projects.

7. Sharing implementation intentions

Imagine a well-functioning academic department that is making a hiring decision. Each faculty member knows different things about the candidates, and each has their own opinion about who to hire. But they also share the goal of hiring the best candidate possible, and they agree on a set of conditions they would like to see satisfied. This group should make a better decision than any individual would make on their own, so long as they share and use the information that's distributed throughout the group. But groups routinely fail to capitalize on their informational advantages; they forget to share crucial information, and when they do share it, it's often ignored unless most group members already know about it (Wieber et al 2012, 278).[10] Fortunately, groups can exploit the same action-guiding techniques as individuals, when their members identify with a shared goal. Where the best alternative is difficult for individuals to see, but identifiable in light of a group's total information, forming an implementation intention ("When we are about to make the final decision, then we will go over the advantages of the non-preferred alternatives again") can increase the likelihood of choosing the best alternative (Thürmer et al 2014). This works because forming such intentions triggers the construction of individual plans to realize part of a shared activity; and this leads group members to behave in ways that bring about the shared end (cf., Bratman 2014).

Something similar happens when groups make temporally extended plans. Consider a group that has a small pot of money to invest. They settle on an initial plan, and occasionally meet to update that plan in light of their successes and failures. This group should be able to revise or abandon their plan if things go badly, so long as relevant information is shared and available. But group members routinely overcommit to ends they have settled upon as a group, and they often overestimate their chances of success (Wieber et al 2012, 278). As with individuals who make similar errors, this is the result of moving from a deliberative to an implementation mindset. And here too, performance can be improved with implementation intentions. When the success of a project declines and requires lowering investments, groups that intend to judge a project as on-lookers, who aren't responsible for earlier decisions, often fail to adjust their plan. As observers, they should be able to disengage; but they don't. By contrast, groups whose members share an implementation intention ("When we are about to make an investment decision, then we will judge the project as

[10] There are many ways for groups to entrench power and hierarchy, and to cause collaboration to collapse in favor of patterns of exploitation and deference. In many groups, polarization effects will arise, and conformity, hierarchy, and prestige biases will often prevent the emergence of egalitarian group structures. I focus, here, on the best-case scenario here, much as Bratman (2014) does. I think that work carried out by Elinor Ostrom (1990) demonstrates that these sorts of cases are possible; but they are difficult to sustain, and the emergence of any pattern of exploitation can prevent a group from capitalizing on its informational advantages.

onlookers who are not responsible for earlier decisions!") adjust their investments to track their current situation. In both cases, people are reflexively led to act as participants in a joint-activity; but those who use implementation intentions can exploit cognitive strategies that allow them to take advantage of the information that is distributed among group members (Wieber et al 2013). By recognizing the potential for epistemic errors, they can use their foreknowledge as group members to guide their joint-activity.

Nonetheless, robust limitations remain. As Dennett and Bratman both argue, planning to do something together allows us to precommit to actions that are consistent with our shared goals and values. And where people think and act together, they can capitalize on shared information to successfully execute joint-actions, so long as they are mutually responsive. But this only works where we already have strong commitments to pursuing a particular goal; like individual intentions, joint-intentions can only link plans to currently salient information, and they can only facilitate top-down control over our behavior. This opens up more degrees of freedom, but it doesn't explain how we can move beyond the thoughts that readily occur to us.

8. Shared plans

This is where things get interesting. I contend that deliberation and planning can be realized by distributed networks of individuals, who use discrete representations to broadcast their own attitudes, and revise them in parallel, in light of shared modes of thinking. This can happen where cooperating individuals convey just enough information to coordinate, using "trading languages" that allow them to query each other while ignoring many irrelevant details (Galison 1997, 883). As a result, acts of planning together become transactive; we rely on information that has been provided by others, and take part in the collaborative construction of prospective representations that can guide our joint-activities. At the same time, this process is guided by subpersonal mechanisms that facilitate alignment in dynamic interpersonal contexts.[11] This is a complex and contentious idea, so let's move through it with a familiar example before turning to the power of this process to create new degrees of freedom.

Imagine a long-term couple that takes a vacation together. Over the years, they have learned that one of them will know where the best coffee is, while the other will know where the best museums are (with substantial redundancy in other domains, and some redundancy even here). When they arrive in a city, one of them might focus on getting to their favorite cafe, while the other might think about heading to a modern art exhibit. Each of them may have an initial inclination to nudge the other, as each will assume that their own plans are most feasible and desirable. But suppose they have also developed strategies for navigating disagreement, and they decide to talk about what to do first. They might realize that getting lunch first will allow them to make further

[11] Thanks are due to John Sutton for pushing me on this point in another context. The details of my argument, here, are articulated more fully in Huebner (2014). But the point runs much deeper than I have time to address in this paper (which is already much too long). Moreover, as Maureen Sie (p.c.) notes, I haven't said much of anything about the role of our moral reactive attitudes in guiding collaborative and collective behavior. She is right to flag this, and any plausible attempt to fill out the ideas that I have sketched in this paper will have to acknowledge the importance of these attitudes to sustaining, undermining, and opening up new possibilities for collaborative action. Finally, Manuel Vargas (p.c.) notes that affect more generally—beyond just the reactive attitudes—is likely to paly a significant role in our attunement to social practices and in the structure of the social scaffolding that we all rely upon. I couldn't agree more (cf., Huebner in press). To be clear, I agree with much of the argument that is developed in Sie (2013) and with Vargas's claim that I must be more attentive to affect, and in my future attempts to expand on this perspective I plan return to these issues in more detail.

decisions without arguing—something neither of them would have noticed before. And by verbally interacting, and attending to patterns of non-verbal behavior, they can construct a shared plan that will account for their independent needs and shared goal of enjoying their vacation. In this respect, their capacities parallel the capacities of couples that can reconstruct shared memories together (Huebner forthcoming; Theiner 2013; Tollefsen et al 2013), and thinking briefly about memory can shed light on the nature of this constructive process.

Like plans, episodic memories are constructed by fleshing out the details of skeletally structured representations (Bartlett 1932; de Brigard 2014; Neisser 1981). We "draw on the elements and gist of the past, and extract, recombine and reassemble them into imaginary events that never occurred in that exact form" (Schacter & Addis 2007, 27). And like plans, the construction of memories often draws on information that is anchored in the material and social structure of our world (Hutchins 2005). Long-term couples can often minimize the demands on limited cognitive resources by storing different kinds of memories (often with substantial redundancy); this allows them to increase the breadth of their knowledge, while simultaneously increasing its depth, using a virtual memory store that spans the transactive network that they constitute (Wegner 1995). Such couples then construct representations of past events by dynamically adjusting and recalibrating their individual mental states to track the explicit representations that they broadcast to one another. As a result, they can often reconstruct more detailed and elaborate memories by cross-cuing one another and engaging in conversation; and their conversations allow them to recall details of past events that neither would have remembered on their own (Harris et al 2011). The process of remembering thus relies on shared computations over linguistic and gestural representations, and the constructive process only needs to be carried out once—in conversation. Transactive memories emerge because couples 'think out loud', and externalize the processes of remembering; by cross-cuing one another, they implement an interfaced system, distributed memories bind the group together, and "any one individual is incomplete without being able to draw on the collective knowledge of the rest of the group" (Wegner & Ward 2013). In essence, they form a flexibly-coupled system that mimics the architecture of the BitTorrent Protocol.

Likewise, when couples plan together, they can retrieve and broadcast individually stored representations, using a process of cross-cuing to construct a shared plan that doesn't need to be represented prior to conversation. By constructing a plan together, they implement an interfaced system that produces shared representations, which guide collaborative behavior, and generate individual expectations that are shared by individuals *qua* group members, just as Bratman suggests. As they flesh out the details of their plan, facts that are crucial to their current interactive context may arise in ways that neither of them would have considered on their own; consequently, they might act in ways that diverge from the preferred option that either would have selected, but that are well suited to their interactive context. Domain-general computations over linguistic and gestural representations are used to construct shared plans, and there is no reason for this process to occur more than once. By 'thinking out loud', people can thereby externalize the deliberative process, yielding explicit representations that are relevant to the guidance of collaborative behavior, even though the implementation of the resulting plans remains an individual process.

We can plan together as group members who share a great deal of epistemic common ground, and we often do so. But this can create epistemic echo chambers, where ideas are replicated, sustained, and more deeply entrenched.[12] The regulative dimensions of mindshaping have precisely this effect. They make it easier for us to accept habitual patterns of thought and action, and easier for us to passively accept the power structures that arise through processes of cultural evolution

[12] I borrow this idea of an epistemic echo chamber from Benjamin Elzinga.

(Dennett in prep). We are inclined to act in accordance with social norms, and as a result we often find ourselves on local peaks in an adaptive landscape, happy enough with our situation, and unable to imagine other ways that the world can be. This isn't always a bad thing. Those of us who benefit from local power structures tend to act in ways that feel right to us. And this shouldn't be a surprise; Darwinian evolution tends to be a stabilizing force, which preserves traits that help animals occupy the available niches. But we can also construct novel niches, because we can imagine possibilities that are better than the ones that we have come to expect.

9. Help from our friends

The forms of joint-deliberation and joint-agency I have just been discussing each play a critical role in the process of imagining another world, and both must be in place to open up and sustain *novel forms of elbowroom*. Joint-deliberation can help us realize that our current ways of thinking have emerged as a result of our contingent learning histories, and it can help us find new ways to think about the world we inhabit. This happens as we express facts about our notional world with explicit representations, and submit them to practices of evaluation and revision; and where shared deliberations arise outside of epistemic echo chambers, when we interact with people who have different expectations and different insights about aspects of our shared practices, this can be a highly productive process. Specifically, when we *listen* to people who have different forms of embodiment than us, or different backgrounds that lead them to expect different things, we can come to see the world in very different ways.

In a strange way, I think that Dennett hit on a similar insight in one of his early attempts to address questions about freedom and responsibility. I paraphrase here, replacing Dennett's (1978, 295) claims about subpersonal processes with claims about supra-personal processes:

> My model of decision-making has the following feature: when we face with a normatively significant decision, we can often rely on others to generate considerations that we wouldn't have access to our own. Which considerations arise will not be fully determined by the fact that we are planning together; and as we evaluate these considerations, some will be immediately rejected as irrelevant, or as inconsistent with our shared goals. The considerations we select as having a reasonable bearing on our shared behavior will then be submitted to practices of giving and asking for reasons, and if we are mutually responsive, and non-exploitative, the considerations we arrive at will ultimately serve as predictors and explicators of our decisions and actions *qua* group members.

It should come as no surprise that this is a common thread that runs through the collaborative practices of early 20th century anarchists, the social justice work of Óscar Romero and Paul Farmer, and the attempts by the Zapatistas of Chiapas to build a better world (Lynd 2012; Lynd & Grubačić 2008). Each group puts *listening* at the core of their radical project. They see that it is often only by embedding ourselves in a world that we don't quite understand that we can begin to change the options that we see, and begin to see the unfounded presuppositions that we have signed onto unreflectively. By attempting to understand the world as others do, we can sometimes recognize the contingency of things that have seemed necessary.[13]

[13] This is one reason why it makes sense to teach all of the worlds religions in schools (Dennett's 2006).

Revealing contingencies, however, is only the first step in breaking down the teleological assumptions that lead people to believe that our social arrangements must continue to be what they currently are; and it is the first step in making more elbow room for new forms of action through acts of planning together. To the extent that we embed ourselves in shared practices that are designed to foster mutual understanding, we may be able to uncover options that we would have missed on our own; and some of these ideas may push us beyond those that would have been available given our reinforcement history. I think Dennett knows this, but I have no idea whether he knows that he knows it (nor even whether he believes it).

Specifically, I contend that a form of mental contrasting can be useful for increasing our freedom. By imagining a desired future, and reflecting on facts that stand in the way of reaching it, individuals can enhance the cognitive relevance of the desirable features of an action, and highlight the feasibility of particular plans (Oettingen & Gollwitzer 2010). But I contend that the reliance of this practice on explicit representations should also make a collaborative form of mental contrasting possible. If so, this should be able to open up an inherently social form of elbowroom. Achieving this, however, is not easy. On a small scale, we only need to reflect on the shared habits that emerge in long-term relationships. Breaking out of these habits often requires interjecting new forms of thinking, and developing new habits of communication, to allow previously undisclosed facts to be brought to the fore. Teaching individuals to do this is big business, and not a particularly successful business at that. It requires cultivating new forms of thinking, as well as abilities to think across different social and cultural frameworks, and the willingness to interact with people from other professional groups, socio-economic status, religions, political persuasions, and more (Gavazzi p.c.). Strategies for doing this will only become more difficult to enact as groups become larger and more diverse.

Nonetheless, there is reason to believe than if individuals become comfortable articulating their conceptions of a desired future, and collaboratively reflect on the facts about the world that stand in their way, they will be able to enhance the cognitive relevance of the desirable features of novel patterns of behavior, and highlight the feasibility of the plans they have constructed as a group.[14] Having articulated a shared idea about which ends to pursue, groups of people can begin to set out plans that will allow them to solidify new ways of living and acting together. Indeed, people often use forms of social-monitoring and self-monitoring to sustain forms of collective action, and they are most likely to succeed where strategies for managing defection and cooperation are self-organized, and grounded in ideals that everyone adopts (Ostrom 1990). Where people agree that an issue they face is important, retain a distributed and collective form of autonomy over mutually agreed upon rules, and develop community-centered practices for monitoring and sanctioning others, they can act to foster ongoing forms of collaboration. Again, this isn't easy, but people can work toward this end by precommitting to particular practices *qua group members*, and thereby decreasing the likelihood of defection.

Of course, we must first find ways to sustain forms of thinking that allow us to act as equals, especially where people dominate discussions or engage in exploitative practices (Bratman 2014). In a world dominated by exclusionary practices, it will be difficult to sustain non-exploitative practices and non-dominating forms of speech. Future-directed intentions can be valuable in this context, as can pre-commitments that impose normative pressure against those who defect from shared

[14] One intriguing piece of data in this regard is reported by Anita Wooley and her colleagues (2010), who found that the performance on a wide range of tasks, for people working in small groups of two to five people, is not predicted by the average or maximum individual intelligence of group members, but is highly correlated with things like the average social sensitivity of group members, equitable turn-taking, and the proportion of women in a group.

practices. As group members, pressure isn't just psychological; it's also social. And by changing our patterns of social engagement, we can begin to pull ourselves toward counter-normative, yet preferable ideals with help from our friends.

Finally, planning together can create novel environmental contingencies, which can prevent the forms of backsliding to which Bayesian agents are susceptible. The structure of the cognitive prosthetics we rely upon has an enormous impact on the ways that we update ongoing behavior. We attune to social practices, and our high-level attitudes and low-level reactions tend to entrain to the local patterns we encounter, leading to stable institutional structures against which future attitudes and behavior attune. By building stable micro-worlds, which accord with our forward-looking expectations, we can get the feedback loop that I discussed above to solidify ways of thinking and acting that are consistent with our values and ideals, instead of allowing the world to undermine them. Collaborative actions that are grounded in prefigurative imagination will be met with evaluative feedback suggesting that we are acting rightly. As a result, new hypotheses can be sustained, which contradict dominant forms of social power; and as the brain searches for the linked set of hypotheses that make the incoming data most plausible, our expectations can begin to shift toward the local patterns of interaction we are in the process of constructing. To my mind, this seems like the kind of freedom that is most worth wanting: it is the political freedom to change social norms, by collaboratively resisting them in the short run, and entrenching them in shared social practices that open up novel degrees of freedom.[15]

10. Epilogue

There is a great deal more work to be done in fleshing out a plausible, socially situated view of agency. And my claims in this chapter are probably too optimistic, given the pervasive role of exclusionary ideals and hierarchical ideologies in the world we inhabit. So for now, I just want to note that we should not be content with the kinds of freedom we can pursue on our own. Since the human brain is a predictive machine, we are constrained by social and environmental contingencies. But the point of doing philosophy isn't simply to understand *that* we are constrained, it is to find ways of changing *how* we are constrained. We are rabid social niche constructors, and another world is possible. But acting in ways that go beyond our contingent learning histories requires planning together, imagining together, and acting together in ways that prefigure the world we would like to inhabit. In the process, our own preferences will probably need to change, and they are only likely to do so as a result of accompanying others in new kinds of shared practices.

When we act on our own, we will be lucky if our behavior happens to align with values and ideals that we can reflectively avow. But this is not guaranteed, even where things feel relatively comfortable. By contrast, where we collaboratively build social structures, our values and judgments

[15] People who live in more diverse communities, and who interact with the members of other racial groups in a more diverse range of situations, for example, tend to be less racially biased, and tend to have explicit attitudes that are more egalitarian (Dasgupta & Rivera 2008). Inhabiting such neighborhoods creates and reinforces positive implicit associations, which can counteract the biases that structure the rest of our world (Dasgupta 2013, 247). Living in such neighborhoods also helps mitigate the appeal of colorblind ideologies, and heightens the awareness of forms of structural racism that go beyond explicitly racist attitudes (Dasgupta 2013; Garner 2007, 45-46). Such attitudes can help us to see the hegemonic ideology of Whiteness as contingent, distorting, and dispensable, instead of seeing it as a necessary conceptual framework. Of course, power is never given up easily, and there are many opportunities to abandon anti-racist attitudes in favor of the comfort of White ideology. But diverse spaces, structured around diverse goals and values, may help to provide a place for anti-racist ideologies and critical approaches to Whiteness to develop. For a slightly more robust discussion of these issues, see Huebner (2015).

can stabilize around positive and productive biases, which will become calcified in social norms and practices, and guide ongoing behavior as free-floating rationales. Put somewhat differently, planning and acting together can yield a form of social niche construction, which is grounded in our capacity to think and act together; as associative thinkers, the ends we seek become more stable as we surround ourselves with others who value the same things as us (Spinoza 1677/2002, 339). The Cartesian approach to 'free will' that tends to dominate discussions of agency focuses attention on ways of changing ourselves, and making ourselves better people. But as naturalist philosophers, we should recognize that we are often nudged around by interactions with the world in ways that we cannot control, but as groups working together, we can collectively devise circumstances that enhance our collective power to act. So the ethical question should always be: How can we construct forms of collective action that open up possibilities that we don't possess as individuals. And this makes ethics, as well as reflections on agency and freedom, a matter of politics, not metaphysics.

11. Acknowledgements

This ideas in this paper developed over the course of several helpful conversations with Gunnar Björnsson, Justin Caouette, Gregg Caruso, Mattia Gallotti, Joey Jebari, Pete Mandik, Manuel Vargas, Tad Zawidzki, and of course Dan Dennett. John Gavazzi, Ruth Kramer, Rebecca Kukla, and Maureen Sie each read a complete draft of the paper, and offered insightful comments on where I had gone wrong. I think it's fair to say that I got by with quite a bit of help from my friends

12. Works Cited

Adriaanse et al (2011a). Do implementation intentions help to eat a healthy diet? Appetite, 56, 183-193.

Adriaanse et al (2011b). Breaking habits with implementation intentions. Personality and social psychology bulletin, 37, 502-513.

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Akins, K. (1996). Of sensory systems and the" aboutness" of mental states. *The Journal of Philosophy*, 337-372.

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Review,* 6, 292-297.

Barrett, L.F. (2015). When a gun is not a gun. New York Times, 17 April 2015; retrieved 12 August 2015 from http://goo.gl/PvXLDt.

Bartlett, F. (1932). Remembering. Cambridge: Cambridge University Press.

Belanger-Gravel, A., G. Godin & S. Amireault (2013). A metaanalytic review of the effect of implementation intentions on physical activity. Health Psychology Review, 7, 23–54.

Blomberg, O. (in press). "Review of Shared Agency", Analysis

Botvinick, M. & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences, 10*, 485-588.

Brandstätter, V., & Frank, E. (2002). Effects of deliberative and implemental mindsets on persistence in goal-directed behavior. *Personality and Social Psychology Bulletin*, *28*(10), 1366-1378.

Bratman, M. (1987). Intention, Plans, and Practical Reason. Cambridge: Harvard University Press.
Bratman, M. (2008). Intention, Belief, Practical, Theoretical. In S. Robertson (ed.), Spheres of Reason (pp. 29-51). Oxford: Oxford University Press
Bratman, M. (2014). Shared Agency: A Planning Theory of Acting Together. New York, Oxford University Press
Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies*, *117*(1), 143-164.
Clark, A. (1996). Linguistic anchors in the sea of thought?. *Pragmatics & Cognition*, *4*(1), 93-103.
Clark, A. (this volume). Strange Inversions: Prediction and the Explanation of Conscious Experience.
Coleman, N.A.T (2015), Gender, Race, and Philosophy, "How Philosophy Was "Whitewashed"? Retrieved on 11 Sept 2015, from: http://goo.gl/cCypbS.
Couzin, I. D. (2009). Collective cognition in animal groups. *Trends in cognitive sciences*, *13*(1), 36-43.
Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.
Crockett, M. et al. (2013). Restricting temptations. *Neuron*, *79*(2), 391-401.
De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155-185.
Dehaene, S., & Changeux, J. P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Progress in brain research*, *126*, 217-229.
Dennett, D.C. (1978). On giving libertarians what they say they want. *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: MIT press.
Dennett, D.C. (1995). Darwin's Dangerous Idea. *New York: Simon and Schuster*
Dennett, D.C. (2003). *Freedom evolves*. London: Penguin UK.
Dennett, D.C. (2006). *Breaking the spell: Religion as a natural phenomenon*. London: Penguin UK.
Dennett, D.C. (2015). Why and How Does Consciousness Seem the Way it Seems? Open MIND: 10(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570245
Dennett D.C. (in prep). From Bacteria to Bach and Back.
Dewey, J. (1930). From Absolutism to Experimentalism. In G. Adams & W. Montague (eds). *Contemporary American Philosophy: Personal Statements*. Russell and Russell: 13-27.
Drayson, Z. & A. Clark (forthcoming). Augumentation, Agency, and the Spreading of the Mental State.
Fitch, W., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*,*1316*(1), 87-104.
Friston, K. (2009). The free-energy principle: a rough guide to the brain?.*Trends in cognitive sciences*, *13*(7), 293-301.

Gallotti, M., & C. Frith (2013). Social cognition in the we-mode. *Trends in cognitive sciences*, *17* (4), 160-165.

Galison, P. (1997). *Image and logic*. Chicago: University of Chicago Press

Gilbert, S. et al (2009). Separable brain systems supporting cued versus self-initiated realization of delayed intentions. Journal of Experimental Psychology. Learning, Memory, and Cognition, 35, 905–915.

Gollwitzer, P. (1999). Implementation intentions. American Psychologist, 54, 493–503.

Gollwitzer, P. (2012). Mindset theory of action phases. In P. VanLange et al (eds.), Handbook of theories of social psychology (Vol.1, pp. 526-545). London: Sage Publications.

Gollwitzer, P. M. (2014). Weakness of the will: Is a quick fix possible? Motivation and Emotion, 38, 305-322.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in Experimental Social Psychology, 38, 69–119.

Gross, Z. (2012). AAPD Interns Blog, Navigation and disabled interdependence. Retrieved on 21 December 2014, from: goo.gl/s8iQai

Harris, C., P. Keil, J. Sutton, A. Barnier, & J. McIlwain (2011) We Remember, We Forget, Discourse Processes 48, 4: 267-303.

Holroyd, J. (2011). The metaphysics of relational autonomy. In C. Witt (ed). Feminist Metaphysics. Dordrecht: Springer, 99-115.

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

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Huebner, B. (2014). Macrocognition. New York: Oxford University Press.

Huebner, B. (2016). Transactive memory reconstructed. Southern Journal of Philosophy.

Huebner, B. (in press). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. *Implicit bias and philosophy: Metaphysics and Epistemology*.

Humphreys, G., E. Forde, & M. Riddoch (2001). "The neuropsychology of everyday actions," in B. Rapp (ed) *The handbook of cognitive neuropsychology*. Cambridge, Mass.: MIT Press, 565–592.

Hutchins, E. (2005). "Material anchors for conceptual blends," Journal of Pragmatics 37, 10: 1555-1577.

Jackendoff, R. (2007). *Language, consciousness, culture: Essays on mental structure*. Cambridge, MIT Press.

Kim, M., Loucks, R., Palmer, A., Brown, A., Solomon, K., Marchante, A., & Whalen, P. (2011). The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. *Behavioural brain research, 223,* 2, 403-410.

Kirsh, D. (2006). Distributed cognition: A methodological note. *Pragmatics & Cognition*, *14* (2), 249-262.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513-549.

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61(1): 140-151.

Klucharev, V., Munneke, M.A., Smidts, A., & Fernandez, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience* 31: 11934-11940.

Kosslyn, S. (2006). "On the evolution of human motivation," in S. Platek, T. Shackelford, & J. Keenan (Eds.), *Evolutionary cognitive neuroscience*. (Cambridge: MIT Press).

Kukla, R. (2005). Conscientious autonomy: displacing decisions in health care. *Hastings Center Report*, *35*(2), 34-44.

Kutz, C. (2000). Acting Together. Philosophy and Phenomenological Research 61 (1): 1-31.
Landsiedel, J., & S.J. Gilbert (2015). Creating external reminders for delayed intentions: Dissociable influence on "task-positive" and "task-negative" brain networks. *NeuroImage*, *104*, 231-240.
Lashley, K.S. (1951). The problem of serial order in behavior. In L.A. Jeffress (Ed.), Cerebral mechanisms in behavior. New York: Wiley.
Lynd, S. (2012). Accompanying: Pathways to social change. PM Press.
Lynd, S. & A. Grubačić (2008). Wobblies and Zapatistas: Conversations on Anarchism, Marxism and Radical History. PM Press.
MacKenzie, C. (2014). Autonomy. In J. Arras, E. Fenton, & R. Kukla, (eds), The Routledge Companion to Bioethics. New York: Routledge,
Mackenzie, C. & N. Stoljar, eds (2000). Relational Autonomy: Feminist Perspectives on Autonomy, Agency and the Social Self. New York: Oxford University Press, 277-290.
Mahon, B. & A. Caramazza (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102* (1), 59-70.
Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes. Personality and Social Psychology Bulletin, 36 (4), 512-523.
McLeod, C. (2002). Self-trust and reproductive autonomy. Cambridge: MIT Press.
Montague, R. (2006). *Why Choose This Book?* New York: Dutton Press.
Neisser, U. (1981). "John Dean's memory," *Cognition*, 9: 1-22.
Oettingen, G., & Gollwitzer, P. (2010). Strategies of setting and implementing goals: Mental contrasting and implementation intentions. In J.E. Maddux et al (Eds.), Social psychological foundations of clinical psychology. New York: Guilford, 114-135.
Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge: Cambridge University Press.
Pacherie, E. (2006). Toward a dynamic theory of intentions. Does consciousness cause behavior, 145-167.
Pacherie, E. (2012). The phenomenology of joint action. In Axel Seemann (ed.). Joint attention: New developments (pp. 343-389). Cambridge: MIT Press
Polanía, R., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. Nature communications, 6.
Rescorla, R.A. (1988). Pavlovian conditioning: It's not what you think it is. American Psychologist, 43(3), 151.
Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & Van Der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. Human movement science, 26(4), 525-554.
Schelling, T. C. (1966). *Arms and Influence*. New Haven, CT: Yale University Press.
Schacter, D. & D. Addis (2007). The cognitive neuroscience of constructive memory, Philosophical Transactions of the Royal Society B, 362, 773-786.
Schwabe, L., & Wolf, O. T. (2013). Stress and multiple memory systems. Trends in cognitive sciences, 17(2), 60-68.

Seligman, M., P. Railton, R. Baumeister, & C. Sripada (2013). Navigating into the future or driven by the past. Perspectives on Psychological Science, 8(2), 119-141.
Shea, N. et al (2014). Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, *18*(4), 186-193.
Sie, M. (2013), Free Will an Illusion? An Answer from a Pragmatic Sentimentalist Point of View, in: Caruso, G. (ed.), *Exploring the Illusion of Free Will and Moral Responsibility,* Lexington Books, Rowman & Littlefield, pp 273-289.
Sie, M. (2014). Self-knowledge and the minimal conditions of responsibility. The Journal of Value Inquiry, 48(2), 271-291.
Simon, H.A. (1969). The sciences of the artificial. Cambridge: MIT Press.
Spinoza, B. (1677/2002). *The Complete Works*. S. Shirley & M. Morgan, eds. New York: Hackett Publishing.
Sparrow, B., Liu, J., & Wegner, D.M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776-778.
Taiwo, O. (in prep). Why We Are Not What We Seem: The Social Ascription Critique of Agency.
Theiner, G. (2013). Transactive Memory Systems: A Mechanistic Analysis of Emergent Group Memory. Review of Philosophy and Psychology, 4(1), 65-89.
Thürmer, J., F. Wieber & P. Gollwitzer (2014). When unshared information is the key. *Journal of Behavioral Decision Making*.
Tollefsen, D. (2005). Let's Pretend: Children and Joint Action. Philosophy of the Social Sciences 35 (1): 75-97.
Tollefsen, D. (2014) A dynamic theory of shared intention. In S. Chant & F. Hindriks (Eds.) *From Individual to Collective Intentionality* (pp. 13-33). Oxford: Oxford University Press.
Tollefsen, D. & R. Dale (2012). Naturalizing Joint Action. Philosophical Psychology, 25 (3): 385 - 407.
Tollefsen, D., R. Dale & A. Paxton (2013). Alignment, Transactive Memory, and Collective Cognitive Systems. Review of Philosophy and Psychology 4 (1):49-64.
Taylor, S. E., & Gollwitzer, P. M. (1995). Effects of mindset on positive illusions. *Journal of Personality and Social Psychology, 69,* 213-226.
Tollefsen, D. 2014. A dynamic theory of shared intention. In S. Chant & F. Hindriks (Eds.) From Individual to Collective Intentionality: New Essays, Oxford University Press: Oxford.
Webb, T., P. Sheeran, & A. Pepper, 2012, Gaining control over responses to implicit attitude tests, British Journal of Social Psychology, 51(1): 13–32.
Wieber, F., J. Thürmer, & P. Gollwitzer (2012). Collective action control by goals and plans. American Journal of Psychology, 125, 275-290.
Wieber, F., J Thürmer, & P. Gollwitzer (2013). Intentional action control in individuals and groups. In G. Seebaß et al (Eds.), Acting intentionally and its limits (pp. 133-162). Berlin: DeGruyter.
Wilson-Mendenhall, C.D., Barrett, L.F., Simmons, W.K., & Barsalou, L.W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia* 49: 1105-1127. Supplemental Materials.
Wegner, D. (1995). "A computer network model of human transactive memory," Social Cognition, 13, 1-21.
Wegner, D. & A. Ward (2013). The Internet Has Become the External Hard Drive for Our Memories. Scientific American, 309, 6. http://goo.gl/vEB1h2 (retrieved 1 February 2015).
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330* (6004), 686-688.