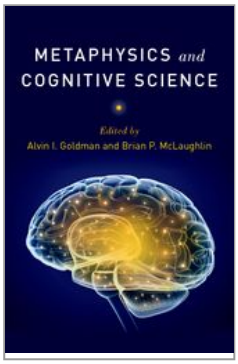

Metaphysics and Cognitive Science
Alvin I. Goldman and Brian P. McLaughlin

Print publication date: 2019
Print ISBN-13: 9780190639679
Published to Oxford Scholarship Online: April 2019
DOI: 10.1093/oso/9780190639679.001.0001

Modal Prospection

John McCoy
L. A. Paul
Tomer Ullman

DOI:10.1093/oso/9780190639679.003.0010

Abstract and Keywords

Drawing together the metaphysics of counterfactuals with empirical work on intuitive judgments, this chapter discusses the nature of counterfactual reasoning about self-involving possibilities. It argues that when a person reasons about her self-involving possibilities, especially far-fetched possibilities, this reasoning may be supported by an underlying “self simulator,” a kind of mental engine with an approximate understanding of who she is, which enables her to learn about her preferences and make intuitive judgments and predictions about her self-involving possibilities. On this view, through observing or simulating their own choices, people understand themselves through a process similar to that by which they understand other people. In this way, they learn about their own beliefs and desires. The argument is informed by empirical data from surveys where lay people and philosophers decided what action they would take in vignettes involving a potentially transformative decision.

Keywords: modality, counterfactual, de se, imagination, transformative experience, possibility, theory of mind, intuitive theory, decision-making, theory of self

9.1 The Metaphysics of Modality Meets Cognitive Science

Recent work in metaphysics has been enriched and inspired by work in foundational physics and the philosophy of physics. We think there is even more potential for productive collaboration between cognitive science and metaphysics.[1] To this end, our chapter develops new connections between

metaphysics and cognitive science, drawing together intuitive modeling and prediction in cognitive science with the metaphysics and epistemology of counterfactual reasoning.

Our focus is on how we interpret, represent, and understand possibilities. We are particularly interested in reasoning about certain sorts of self-involving possibilities, especially far-fetched possibilities for oneself or for individuals we are close to. However, our work connects to a wide variety of more general topics, such as those involving modal reasoning, simulation theory, the semantics of counterfactuals, modes of presentation, conceivability and possibility, decision-making, and debates about the nature of perspectival, or *de se*, thought and content.

This first section of our chapter develops a theoretical and empirical context for assessing certain kinds of possibilities that brings together metaphysics, epistemology, and computational cognitive science, and discusses ways to connect this richer perspective to other topics in philosophy, such as the concept of transformative experience, the semantics of counterfactuals, moral learning, and simulation theories. We start by drawing connections between counterfactual reasoning about physical possibilities and recent empirical work on intuitive physics judgments. We then extend this idea to reasoning about other people and to counterfactual reasoning about self-involving possibilities, and explore the parallels between intuitive self-involving judgments and intuitive physics judgments. We take the view that, when a person reasons about her self-involving possibilities, especially far-fetched possibilities, this reasoning may be supported by an underlying "self simulator," a kind of mental engine with an approximate understanding of who she is, **(p.236)** which enables her to learn about her preferences and make intuitive judgments and predictions about her self-involving possibilities. We then introduce the notion of modal prospection, and discuss connections between the ideas we are exploring and some contemporary philosophical debates in metaphysics, mind, and epistemology. In section 9.2, we consider two sample vignettes in which participants are asked to make a potentially transformative decision, and explore the philosophical and methodological reasoning behind our surveys, with special attention to our use of far-fetched, fantastical examples, the stock-in-trade of the metaphysician. Our farfetched possibilities involve metaphysically possible scenarios that highlight meaningful or fundamental elements of how people think about themselves, and we explain why we think these scenarios are especially apt for our purposes.[2] Section 9.3 presents the empirical part of our project, and discusses the ten self-involving possibilities in the surveys we conducted. We present the details of our surveys, and discuss the empirical and theoretical questions that devolve from our results, laying the groundwork for further work on this rich interdisciplinary topic. Our appendix lists the vignettes we used.

9.1.1 Background I: The Metaphysics of Counterfactuals and the Cognitive Science of Intuitive Physics Judgments

On a standard approach (Lewis 1986; Stalnaker 1976), reasoning about possibilities involves reasoning about possible worlds. If Finbarr drops a glass on the sidewalk, it will (likely) break: at time t in a world with laws just like the actual world, just like our world but for the initial state at t in which Finbarr drops his glass, it falls to the sidewalk and breaks. The metaphysics of counterfactuals, when understood using this possible worlds framework, involves the notion of similarity: we assess various modal claims in virtue of similarity relations between worlds. Very roughly, to evaluate counterfactuals such as "if C had not occurred, then E would not have occurred," we move to the closest possible world where C does not occur. If E does not occur in that world, the counterfactual is true. Closeness of worlds, here, is based on relevance and similarity: the closest possible world is the world that is most similar to the world of evaluation (usually the actual world) in salient respects. More generally, to determine whether S is possible in world W, we need to know which possible worlds are most similar to W in the relevant ways. Different similarity relations will define different kinds of possibilities.

Cognitive scientists are interested in how the mind represents the world when people make fast, natural, and intuitive judgments about ordinary goings-on in their environment.[3] How am I representing the world when I quickly, intuitively, **(p.237)** and naturally judge that, if Finbarr were to drop the glass on the sidewalk, it would (likely) break? Recent research on intuitive physics judgments (e.g. Gerstenberg et al. 2012; Battaglia et al. 2013; Sanborn et al. 2013; Smith and Vul 2013; Hamrick et al. 2016; Ullman et al. 2018) frames this sort of understanding of the dynamics of the world as people having a *world simulator* in their heads where this world simulator is a "mental physics engine," in analogy to the physics engines software that powers modern animations and computer games (although this is not the only way people may be making intuitive physics judgments; see e.g. Forbus 1988; Marcus and Davis 2013). A physics engine is software that generates a simulation of a dynamic physical system, such as simulations of collisions, explosions, or the movement of fluids. In particular, the analogy is to physics engines that support relatively fast and approximate simulations like those that power games, rather than engines underlying scientific simulations such as galaxy formation or protein folding (Ullman et al. 2017). As well as using their mental physics engine to evolve the world forward and predict what will happen, people can also use it to reason about the physical properties of a scene. For example, by observing objects collide, people can make inferences about the mass of the objects or the friction of the surface on which they were moving.

According to this account, the mind can use a quasi-Newtonian simulation to predict how a physical scene would unfold over a short time-span, in the same way that a real physics engine can quickly simulate the results of acting in a

game world. My intuitive, commonsensical judgment that if Finbarr drops his glass it will probably break is based on the underlying computations of a mental physics engine with an approximate understanding of bodies and the forces acting on them.

We can now see a structural parallel between the metaphysics and the cognitive science. From the metaphysics perspective, we can consider a counterfactual claim about some ordinary goings-on in the local environment: “If Finbarr were to drop the glass, it would break.” To evaluate this counterfactual, we move to the closest possible world where Finbarr drops the glass (and all other relevant features are the same), and evolve that world forward. If the glass breaks in that world, then we judge the counterfactual as true. If the glass doesn’t break, then we judge the counterfactual as false. Assume, in the actual world, that the counterfactual is true, and I know that it is true. Then I know something about the nature of this part of my environment; for example, I know something about the laws of this world, and I know something about the disposition of Finbarr’s glass to break. I know that the world is such that the laws make these sorts of ordinary counterfactuals true. But since my knowledge didn’t require any sort of sophisticated physics knowledge, we can also say that part of what I know concerns manifest physics, the physics of ordinary goings-on at the level of human experience.

The manifest image concerns the world as it appears to us. Manifest physics can be thought of as concerning manifest laws of nature, that is, as concerning an implicit representation of what’s generating events like the breaking of the glass. We can think of it in terms of a representation of a machine running the appearances, **(p.238)** a machine that is implicitly represented by us when we understand and predict the ordinary goings-on in our environment. If I can successfully assert and know ordinary counterfactuals like “If Finbarr were to drop the glass, it would break,” then I understand enough about the manifest physics of the actual world to make such predictions. Importantly, this means I understand enough about which features of the actual world, including its manifest features, I need to hold fixed when I determine which possible world is closest, as I assess the truth value of the counterfactual. Note that while we’ve framed our discussion in terms of the manifest, our account can support a realist approach to counterfactuals. The fact that our simulations are largely successful can be taken as evidence that we are getting something right about the nature of the world: we are in fact grasping counterfactual truths.[4]

Blending the cognitive science with the metaphysics, we can think of my counterfactual claim (“If Finbarr were to drop the glass, it would break”) as a judgment based on my underlying computation involving a quasi-Newtonian simulation derived from my approximate understanding of the appearances involving the dynamics, that is, from my approximate understanding of the manifest physics. This simulation is, in effect, a representation of the possible

world W1 where Finbarr drops the glass and it breaks: in some (admittedly implicit sense) the simulation evolves W1 forward from the dropping of the glass to the breaking of the glass. If my simulation correctly predicts the approximate result (the breaking of the glass), then we can interpret this in terms of my selecting the closest possible world, and by extension, correctly identifying the relevant similarity relation: the world I chose for my simulation (W1) was in fact the closest possible world in the relevant respects.

9.1.2 Background II: Indexicals, Self-Locating Attitudes, and Theory of Mind

Philosophers have discovered that reasoning about certain kinds of possibilities can require a more fine-grained modal semantics. In particular, reasoning about possibilities targeted to particular individuals, places, and times requires a semantics that can handle these types of targeted possibilities. Revising our sample counterfactual to include more indexicality, I can assert: "If I were to drop this glass right now, on this sidewalk, it would break." To evaluate the truth value of this counterfactual, I need to assess a possible world where I (or my counterpart) drop my glass at the designated time and place. One popular approach for such assessment uses a semantics that distinguishes the context of utterance from the circumstance of evaluation (Kaplan 1989) in order to determine truth values for claims that are sensitive to indexical content. Our interest in indexical claims can be expanded to self-locating claims, such as "I am the person spilling the sugar," and beliefs based on this, such **(p.239)** as my believing that "I am spilling the sugar." Some argue that reasoning about self-locating (or *de se*) attitudes requires a semantics involving centered worlds, where, to determine the content of the attitude, in addition to specifying the world of evaluation, we must also specify the relevant individual and time. Whether we endorse a semantics that includes *de se* content, or endorse this sort of centered worlds reasoning, or even whether we endorse a particular treatment of indexicality, it seems clear that indexical, self-involving judgments require a distinctive semantics, and this extends to indexical, self-involving counterfactual judgments.

We can make a related claim with respect to the cognitive science about the need for a distinctive model for intuitive, targeted judgments of particular types. One such relevant class of judgments involves judgments about the minds and actions of people. Just as psychologists have studied how people represent the world and simulate it forward, psychologists have also studied how people represent the minds of other people, and use this representation to simulate the actions of other people, given their beliefs and desires. These *intuitive psychology* judgments happen quickly, automatically, and consistently.[5] One influential account of the representations underlying these judgments hypothesizes that these representation takes the form of a "theory of mind" about other people (Wellman and Gelman 1992; Dennett 1987). According to this view, people assume that other people have minds, which include hidden

unobservable mental variables (such as beliefs and desires), which cause observable actions.[6]

On our version of this approach, when I intuitively make the prediction that Finbarr will go to the movie theater today, we frame my understanding of the dynamics of Finbarr's mind as involving a planning algorithm, which takes in mental variables (such as Finbarr's beliefs about the particular movies that are playing today, Finbarr's relative preference for movies over opera, and so on), and generates the next likely action. To capture this idea, we'll describe people as having *agent simulators*. When I predict what Finbarr will do, I employ my agent simulator to approximately predict the likely actions of other people. The notion of an agent simulator is akin to the world simulator of intuitive physics: the way we understand others using a mental agent simulator is like the way we understand the manifest world using a mental physics engine. An approximate world simulator can be used to reason from observed data to hidden dynamic variables (such as mass and friction) or to predict a subsequent state of the world; analogously, the agent simulator can be used to learn about other people's hidden mental variables (such as their desires and beliefs) by observing their actions or to predict their subsequent actions. Mental agent simulators have been the topic of recent influential research in theory **(p.240)** of mind, including quantitative action prediction and understanding in adults and children (e.g. Baker et al. 2009, 2014, 2017; Jara-Ettinger et al. 2016).

Just as my intuitive, commonsensical judgment that "if Finbarr drops the glass it will probably break" is based on the underlying computations of a mental physics engine (my world simulator) with an approximate understanding of bodies and the forces acting on them, my intuitive commonsensical judgment that Finbarr will go to the movies today is based on the underlying computations of my mental planning engine (my agent simulator) with an approximate understanding of agents' beliefs, desires, and planning—perhaps more specifically my approximate understanding of Finbarr's beliefs, desires, and planning architecture.

Before we develop our argument further, an important note on terminology is in order: In cognitive science, theory of mind has often been contrasted with accounts that use the term "simulation." Both accounts concern the explanation and prediction of actions, on the part of both other people and ourselves. As mentioned, theory of mind sees people as constructing intuitive theories of themselves and others, positing indirect hidden variables such as belief and desire. Simulation accounts (e.g., Goldman 1989, 2006), by contrast, see people as having direct access to their decision-making apparatus.[7] On this account, people can explain and predict the actions of others by running a direct simulation of those people on the same neural and cognitive machinery they use to make their own decisions (Nichols and Stich 2003). However, theory of mind is also used to predict the actions of others in a way that is well described by the

term *simulation* (in the sense of evolving the world forward in time from a given initial state). A better distinction between the accounts for our purposes would be to say that one account involves direct access **(p.241)** to decision-making mechanisms, while the other account involves intuitive theories (Saxe 2009). Here, we are primarily concerned with intuitive theories, and so the term *simulation* should be understood as referring to evolving the world forward based on such approximate theories, whether in the physical or psychological domain.

Once we have this model in place, we can think about reasoning involving the sorts of self-involving possibilities and preferences that we are interested in. Just as I can reason about what would happen if I dropped the glass, using my world simulator, I can reason about what you might do if you were faced with a particular choice, using my understanding of your mind given by my agent simulator. But there's a further kind of judgment that I can make, a self-involving one. I can reason about what *I* might do if *I* were faced with the same choice. We will explore the possibility that I use something akin to my agent simulator to reason about myself as well as other agents. We'll refer to this as a *self simulator*.

Recall how believing that "if Finbarr were to drop the glass, it would break" can be interpreted as based on a judgment based on my underlying computation involving a quasi-Newtonian simulation, where the simulation is a representation of the possible world W1 where Finbarr (or his counterpart) drops the glass and it breaks. In a parallel fashion, we can treat my self-involving belief that "if I were given the choice, I would pick the red pill" as a judgment based on my underlying computation involving a self simulation, where the simulation is a representation of the possible world W2 in which I am given the choice and I pick the red pill. This self simulation won't necessarily activate the actual decision-making apparatus that would lead me to a given choice, were I actually faced with taking a red or blue pill. Rather, it's a prediction based on my beliefs about my own values and planning apparatus. Our approach investigates a range of possibilities for how simulating others (agent simulators) could relate to simulating oneself (self simulators) and the world (world simulators).

We can extend this to our philosophical treatment of modality, taking the simulations to, in effect, encode a decision about which world to move to. That is, the nature of the simulation generated by the engine of the mind can be interpreted, philosophically, as an implicit specification of the relevant similarity relation. It can be understood as encoding an implicit specification of which ways of understanding one's self matter (and which don't), which in turn define the relevant respects of the similarity relation, the relation that is used to identify the properties and ways of understanding oneself in the relevant possible world. In the semantics of counterfactuals, the properties that matter determine the relevant world for the assessment of the counterfactual. In our

example with the red pill, the relevant properties define a similarity relation that takes W2 to be the closest possible world for the assessment of the counterfactual. Since in W2 my counterpart chooses the red pill, if W2 is indeed the closest world, the claim “If I were given the choice, I would pick the red pill” is true.

(p.242) 9.1.3 Reasoning about Self-Involving Possibilities

The way we’ve blended a philosophical account of reasoning about oneself with an empirical model of self simulation gives us a framework for exploring philosophical questions about modal reasoning from an empirical perspective, in particular, a framework for exploring the way we assess self-involving possibilities.

In the work we present and discuss below, we explore empirical results collected from participants who were asked to imagine and evaluate a series of life-defining, self-involving possibilities. We surveyed a wide range of individuals, including philosophers, and asked them to consider big decisions involving life-changing possibilities of various kinds. Our scenarios involved possibilities such as becoming a vampire, exploring the universe with aliens, freezing time, knowing the future, consulting all-knowing oracles, and other fantastical situations. (See the appendix for our ten vignettes.)

In presenting people with these vignettes, our primary interest was in exploring the way that individuals reflect on and decide about whether to undergo various sorts of transformative experiences (Paul 2014). These kinds of decisions are focused on self-involving possibilities, possibilities concerning significant changes in one’s self. (Of course, such possibilities also concern others, especially the selves of others, and while we don’t focus on this element, we take it to be implicitly represented in most of our discussion, and we discuss it explicitly in places later in the chapter.) When we consider these kinds of self-involving possibilities, we think of what it would be like to be in various hypothetical situations. As we will discuss subsequently, this is an important way to learn about ourselves, and then use what we’ve learned to decide how to act. This approach to decision-making is based on how we imaginatively represent or simulate ourselves in different scenarios, which the metaphysician can understand as representing oneself at future times in different possible worlds. As we’ll define it, *temporal prospection* is the act of representing or assessing one’s own future (or present, or past) experiences. *Modal prospection* is the act of representing or assessing one’s own possible experiences. Such prospection, more generally, is an act involving the assessment of various sorts of self-involving possibilities.

Why is it philosophically interesting to explore the nature of modal prospection from an empirically informed perspective?[8] One important reason stems from the importance of understanding the way we intuitively evaluate such self-

involving possibilities in order to gain insight into how we reason and learn about ourselves.

As we discussed previously, work on theory of mind suggests that we learn about others from the decisions they make. We think this point extends to ourselves: we might think we have some sort of privileged insight into our own preferences, **(p.243)** but in fact, appearances here are often deceptive. In the sorts of vignettes that our participants consider, we don't think the empirical research supports the view that the best way, or the ordinary way, or even the most natural way, for people to approach these types of decisions is through having direct access to the relevant mental states. That is, we don't think the most natural approach to our ordinary decision-making process for self-involving possibilities uses a formal or deductive reasoning process where we start with our preferences and reason our way through a series of steps to determine how to act.

In support of this point, much evidence from psychology shows that we know less about our own minds or reasons for acting than we normally assume (Epley 2014; Wegner 2002). For example, in one of the best-known experiments in social psychology, participants chose between four stockings that were actually identical (Nisbett and Wilson 1977). The later an item was considered, the more likely a participant was to choose it. However, none of the eighty reasons people gave for their choice mentioned the order of viewing, and when explicitly asked whether order could have affected their judgment, only one participant responded affirmatively. People have all kinds of false beliefs about themselves, from overestimating their positive mental and personality attributes relative to others (Kruger and Dunning 1999), to incorrectly picking out an image of themselves merged with an attractive target as their own face (Epley and Whitchurch 2008). Order and framing effects also seem to influence the moral judgments of people with academic expertise and professional training, just as much as they influence the general population (Schwitzgebel and Cushman 2012, 2015).

Given the evidence that we have limited access to our own preferences, we prefer an alternative where we learn about our own preferences from observing or recalling how we responded in the past to that type of situation, or, if we lack the relevant experience, through simulating or imagining the responses we would have in this type of situation. That is, we may take a certain action, or imagine taking the action, and then think, "Why did I do that? I guess I like X more than I thought." We are not claiming that we have less direct access to our own mental states than we have to the mental states of others. Rather, in both cases, it can be important to infer information about underlying beliefs and desires by observing what decision is made.

Our approach to intuitive judgments suggests that, if we want to decide how to act in a novel situation, we may start by imaginatively assessing ourselves in one hypothetical future as opposed to another, and use that to understand the value and desirability of these different possibilities for ourselves. A very natural way (and perhaps, the most natural way) for us to approach a novel decision is to start by simulating ourselves in the proposed scenario(s), seeing how we'd respond, and reverse-engineering our preferences from this response.[9] We don't have to start by **(p.244)** attempting to identify or list our preferences and then deciding how to act. Indeed, in the survey we present in detail subsequently, 75% of people sampled from the standard US population and 53% of the philosophers who took our survey indicated that they had learned something about themselves as a result of taking the survey, which suggests that they had discovered something about their preferences by thinking through the novel scenarios. People predominantly reported learning something about their personality, their current satisfaction with their life, and their attitudes towards family, friends, and relations.

Exploring how we respond in various novel scenarios, then, can be an important way of discovering what we value, and thus of discovering various truths about ourselves.[10] That is, it can be an important way to discover our preferences when faced with a choice between these possibilities. Once we know our preferences, we can decide how to act.

We see many connections to questions of philosophical interest. One obvious connection is with the debate about the nature of decision-making in cases involving transformative experiences (e.g., Pettigrew 2015; Dougherty et al. 2015; Harman 2015). In that debate, a key issue concerns the epistemically transformative nature of the experience under consideration. If an individual lacks the ability to perform an *informed* simulation, she may lack the ability to determine the relevant similarity relation when assessing an important counterfactual. This can mean that she is impeded in her ability to learn about or form her preferences in the way she needs or wants to, with implications for how she is to approach life-defining transformative choices. See Paul 2018 for an account of how these ideas draw together questions about preference formation, perceptual modes of presentation, and the role of experience in determining salient *de se* truths about oneself.

Understanding the evaluation of self-involving counterfactuals under our preferred framework raises additional interesting questions about the metaphysics of mind. If we can understand self-involving judgments computationally in terms of people relying on their mental self simulators, we can ask: what sort of self simulator is the individual using to make her intuitive judgment? What sort of engine is she using to reason about herself? When making intuitive physics judgments, there are—at least conceivably—many different types of physics engines a philosopher could imagine the mind using,

including ones that don't simulate the specific reality people inhabit. A physics engine may be used to simulate worlds with different laws of nature, such as worlds of different dimensions, worlds with odd time-dependent forces, without gravity, and so on. In a similar manner, there are, at least conceivably, **(p.245)** many different types of self simulators we could use to make predictions about what we would think and choose in various scenarios.[11]

That is, the question about the metaphysics of self simulators is analogous to a metaphysical question about laws of nature, in particular, it is analogous to asking, "What are the laws of nature that govern the evolution of the world from t_1 to t_2?" If different kinds of people use different kinds of self simulators when assessing self-involving counterfactuals, they may be, intuitively speaking, using different kinds of mental machines to generate their intuitive judgments. Recall that we can (metaphorically) characterize a self simulator as a machine that your mind uses to generate intuitive judgments about yourself. We are here asking about the nature of this machine: what kind of machine is it? That is, what kinds of simulations does it generate? Just as the mental physics engine in adults generates noisy Newtonian simulations with forces of gravity and collision, rather than, say, noiseless simulations with random time-dependent forces, a mental engine for generating self simulations could generate one kind of simulation rather than another.[12] Our hope is that, just as work in cognitive science supports the thesis that our intuitive physics judgments are grounded by noisy Newtonian simulations and rules out, say, noiseless simulations or other ways of evolving the world forward, empirical work may guide us to a clearer understanding of the actual types of self simulations we use to make intuitive judgments and learn about ourselves.

Speculating further, in the context of our framework, we see several different (nonexclusive) possible models for how we might simulate self-involving possibilities. We might simulate (1) agentially, by modeling the evolution of a (centered) possible world from our own first-personal perspective, that is, from the first-personal, conscious perspective of the individual who is the indexical center. We might describe this as using a subjective mode of presentation when we simulate. Metaphorically, as we simulate, we "occupy our own shoes." Or (2), we could simulate observationally, by modeling the evolution of a centered world using an impersonal or third-person perspective. On this approach, we simulate using a more "objective" epistemic perspective, something akin to taking a "bird's-eye" view on ourselves in a situation. Metaphorically, we simulate as though we were watching ourselves respond in the counterfactual scenario. (For more on the contrast between agential and observational perspectives on oneself, see Pronin and Ross 2006; Paul 2016, 2018.) Other, quite different, models are also possible. For example, we **(p.246)** might (3) simply judge these counterfactuals retrospectively, in a "model-free" sense (Crockett 2013). Further research is needed to disentangle these possibilities and explore whether

different kinds of simulations lead to different judgments about the relevant counterfactuals.[13]

Another philosophical issue concerns the way we reason about others. Prospection can be something we try to do for others when we try to determine what others might prefer in various situations. Such information is important for projects in decision theory, practical reasoning, ethics, and medical ethics, among other topics. (For example, see Bykvist 2006; Carel et al. 2016; Pettigrew 2018; Shupe 2016; Barnes 2015; Briggs 2015; Collins 2015; Dougherty et al. 2015.)

More generally, thinking about the different ways we can make intuitive judgments about ourselves, and understanding more about how we discover and evaluate our preferences and values, should inform philosophical work on metaphysical theories of the self as well as epistemological theories of how we reason about possibilities and understand self-locating attitudes.[14] Understanding the way we self-simulate might be especially fruitful when connected to work on narrative theories of selves and personal identity (e.g., Schechtman 2011; Strawson 2017; Parfit 1984). A better understanding of the psychology of self-involving judgments could also inform current debates about whether we need a distinctive notion of first-personal or *de se* content (e.g., Cappelen and Dever 2013; Magidor 2014),[15] could help us to understand the structure of indexicality in first-personal thought (Recanati 2012) and whether our understanding of the world and the nature of our thought is fundamentally perspectival (McGinn 1983), can enrich our understanding of the relationship of first-personal thought to motivation and action (Perry 1979), and may even be important for various topics in formal epistemology.[16]

(p.247) We can expand on the connections to ethics and practical reasoning by considering a related philosophical project. Railton (2017) explores the idea that prospection is a natural outgrowth of the Humean insight that we understand the world through a combination of experience and cognitive projection.[17] Railton describes how, just as we can project causal structure onto a sequence of actual events, we can project modal structure onto hypothetical events, mentally extending reality in various ways in order to understand possibility. Railton suggests that imaginative projection can inform moral learning via moral prospection and the empathetic assessment of others.[18] We agree, and take the possibilities here to be expansive: imaginative projection can inform a wide variety of prospective assessments, especially, for our purposes, modal prospection concerning transformative changes in one’s own self and others.[19] Prospection and simulation (or imaginative projection) informs learning about and empathy for others and ourselves, with moral learning and morally informed empathy as a special case. In his paper, Railton argues that imaginative prospection teaches us to evaluate moral possibilities from a “non-egocentric” perspective. Again, we take the possibilities here to be expansive. Imaginative

prospection may be understood in more than one way. Agents may evaluate possibilities from a variety of perspectives, egocentric and non-egocentric, and different kinds of evaluations may lead to different assessments of the possibilities.[20]

Finally, we hope it is clear that our chapter provides a more general foundation for understanding new connections between topics in metaphysics and mind and cognitive science. For example, metaphysical discussions involving modality often concern our judgments about what is necessary, possible, and impossible, and the nature and structure of our modal intuitions play a key role in these discussions. An important debate in the metaphysics of mind explores the relationship between metaphysical possibility and various sorts of thought experiments designed to show **(p.248)** that S is indeed possible in the relevant sense. Explorations of whether S is metaphysically possible often involve questions about whether S is conceivable, whether S is imaginable, what the relationship is between conceivability and imaginability, and whether conceivability entails metaphysical possibility. Some have argued that the way in which we imaginatively represent possibilities needs to be clearly and critically assessed in order to determine whether the structure of what we seem to conceive implies the type of possibility we seem to be discovering (Hill 1997; Hill and McLaughlin 1999). Others argue for a range of ways that imagination relates to assessments of possibilities (e.g., Gendler 2010; Nanay 2016; Ninan 2016, Kind 2016; Williamson 2016). The connections our work draws between empirical research concerning the details of the way the mind predictively represents possibilities and the evaluation of self-involving counterfactuals should be of significant interest to those who explore the relationships between possibility and imaginability. Finally, as we noted previously, there are important debates over the nature of simulation theory that span philosophy of mind and cognitive science (e.g., Goldman 2006; Carruthers 2006; Nichols and Stich 2003). Our work explicitly ties together theories of simulation with work in the metaphysics and epistemology of transformative experience, counterfactual semantics, and *de se* reasoning. This opens new avenues for enquiry and collaboration about the nature of simulation, showing how a topic that is central to the philosophy of mind may also be central to debates in metaphysics and epistemology.

The case is clear. Our philosophical understanding of self-involving possibilities, particularly in the context of modal prospection about transformative decision-making, is enriched by engaging with the relevant work in cognitive science. We turn now to a discussion of the empirical part of our project. In the next section, we introduce the sorts of self-involving possibilities we asked people to make choices about, and discuss the features of these decision tasks that we take to be especially probative.

9.2 Modal Prospection

Consider the following scenario.

> Imagine that aliens come down to Earth, and give you the option to go with them on their travels throughout the universe. The aliens are friendly and honest, and tell you that you would see amazing things on your travels with them if you decide to go with them.
>
> If you decide to go, you will have a week to say goodbye to your family and friends. Once you leave, you will never again return to Earth, nor be able to communicate with people on Earth.

Do you go?

What percentage of other people do you think will choose to go?

(p.249) Now consider another scenario.

> Imagine that there is a magical hourglass. If you flip the hourglass, the following happens:
>
> Every person on Earth stops moving, but you are free to move around as you please. You do not age during this frozen time, but you can be hurt and you can die. For example, if you jump off a tall cliff, you will die. The internet, electricity, and so on carry on working. You will remember everything you did during the frozen time.
>
> If you decide to flip the hourglass, you have to decide in advance how long to freeze time for. You cannot change your mind and unfreeze time in the middle.
>
> If you wish to flip the hourglass it must be done now—you will not be able to flip it at a later time.

Would you like to flip the hourglass? For how long?

What percentage of other people do you think will choose to flip the hourglass?

These kinds of questions ask you to perform a distinctive sort of task: they ask you to deliberate about what you might do in a merely possible scenario involving a transformative experience.[21] Previously, we defined the assessment of one's possible experiences as *modal prospection*. Our interest is in "big decisions," that is, in tasks involving modal prospection about life-defining choices. But our scenarios are highly artificial. Why have we chosen such far-fetched, bizarre scenarios, rather than more everyday sorts of scenarios with transformative experiences such as choosing to have a child or taking a job in a foreign country?

First, because we want to force our participants to imagine and reflect on the situation in order to decide what they prefer. That is, we want participants to think through the situation, ideally through simulating themselves in it, and so we used epistemically new, imaginatively far-flung scenarios, including scenarios that are not physically possible (for example, scenarios where you could do things like freeze time). A familiar scenario could allow a participant to respond simply by drawing on previously stored responses to similar situations, or by using a pre-established convention or known scientific fact. While it's true that our participants may have had late-night dorm room conversations about vampires or debates in the bar about whether the government is covering up the aliens creating crop circles, we expect that, prior to our survey, most of our participants had never considered the explicit, self-involving possibilities we raise.[22]

(p.250) Why is it important to have our participants reason about what they would actually do in each scenario? Because our project focuses on the way agents simulate themselves, and we are interested in gathering data on this feature of our mental lives. We are interested in knowing more about this feature of our mental lives because it is psychologically, philosophically, and practically important. Philosophers and psychologists care about the nature of selves, and about discovering fundamental human values, and understanding self simulators may give us knowledge of the nature and structure of selves and human values. More generally, Kappes and Morewedge (2016) argue that mental simulation can substitute for experience with respect to its evidentiary value and benefits for practice. In situations where we lack the relevant experience, such as novel or transformative decision contexts, simulation may play an especially important role. We think that gaining a better understanding of how we simulate our self in potentially transformative futures is relevant to getting a deeper understanding of agent deliberation and choice that is applicable in a wide range of practical and theoretical contexts.

The second reason we use fantastical thought experiments is that we want to isolate key properties of the self, and our thought experiments give us an excellent way to do this. That is, we are interested in identifying properties that are fundamental to the way people think of themselves and the way they live their lives. These scenarios, if they were real, would bring about life-defining changes: if you were actually to leave with the aliens or freeze time, your life would change in a dramatic way. Speaking philosophically, we are interested in properties and values that are fundamental to one's self conception, perhaps even metaphysically fundamental to one's self, or at least, we are interested in properties and values that determine central or core features of one's lived experience and self-understanding.[23]

Another reason we use fantastical examples is that we want to isolate particular properties or values for our participants to consider (the “key” properties and values we describe in the previous paragraph). To do this, we need to abstract away from the irrelevant and potentially distracting features of ordinary, familiar scenarios. We also need to idealize in order to properly isolate the relevant concept. The need for abstraction and idealization to isolate the concept of interest is familiar from work on the theory and practice of scientific theorizing and discovery. Consider the important role of the Maxwell’s Demon thought experiment in the development of our understanding of thermodynamics, and its continued relevance today in teaching and learning about entropy.[24]

Our fantastical examples, then, are designed to isolate particularly important properties and values, and to highlight what we really care about when making high-stakes, forced choices between contrasting values. For example, the vignette **(p.251)** involving friendly aliens offers you a stark choice: which do you value more, discovery and novelty, or your attachments here on Earth? Our magical hourglass vignette frames a different type of question: would you be willing to isolate yourself from the rest of humanity to gain freedom from the relentless pace of everyday life? To have time to explore and reflect, uninterrupted? Reflecting on what you’d do in these cases is a way of thinking imaginatively, yet precisely, about who you are and what you care about. Would you choose to go with the aliens? Or would you elect to stay on Earth? Would you freeze time, in the process freezing all of your family and friends, and temporarily cutting yourself off from all human interaction? If so, for how long?

So while these situations are far-flung and removed from reality, they connect back to things that many people care about. Their underlying structure concerns deep and fundamental questions about who we are, the sorts of questions we grapple with when making decisions to undergo big life changes. Metaphysically speaking, reflecting on these cases is a way of thinking about the person or self that you are in terms of the values and desires that structure your preferences and define your psychological profile. Arguably, if you’d choose to go with the aliens, you are someone who values exploration and discovery over the status quo. If you choose to stay on Earth, you might be someone who values family and friends over discovery of the unknown. If you choose to freeze time, you might value the freedom this gives you over having contact with friends and loved ones and the rest of humanity, and the “duration” of the freezing you choose may reflect something further about your values.

An additional benefit of our fantastical contexts is that they showcase the particular challenges people can face when they find themselves in high-stakes, highly unusual situations. In such situations, even in the real world, we can find ourselves effectively without guidance, with little anecdotal or scientific information available. Compare the way we’d use our mental physics engine in a high-stakes, novel, but potentially real-world situation where we lacked

background experience or detailed scientific guidance. Imagine being marooned on a desert island, where your only chance of survival was to build a seaworthy boat out of some cardboard that had washed up on the beach. On the assumption that you've never had to do this before, and that you lack detailed knowledge of the engineering required to build such a boat, your best option is to rely on your intuitive physics judgments in order to construct your craft. Similarly, in a high-stakes, novel situation involving a big life decision, your best option may be to rely on your simulations when making your decision, for they may be your only guide. (Alas, for most of us, our intuitive understanding of physics is probably as bad at helping us construct a seaworthy cardboard boat as our intuitive understanding of ourselves is at helping us make good decisions in novel situations. And yet, in some circumstances, it may be all we have.)

We can now address a final issue. Subjectively, to us (the authors), it feels like the scenarios are capturing something important when we deliberate. Why? Why does **(p.252)** it seem so relevant, and important, and interesting to imagine these kinds of highly speculative, fantastical scenarios, as opposed to reflecting on more mundane, ordinary cases? Why does reflecting on these types of decisions seem so meaningful?

The answer can be drawn out from what we've established already.

We care about understanding ourselves and our values. However, as we've suggested, it may be impossible to divine these fundamental values merely by introspecting and trying to perceive them directly. We may have no more access to these mental constructs than we do to the mental constructs of other people, for whom we construct a theory of mind (Saxe 2009). Our fantastical thought experiments give us another way to discover ourselves. They engage the theory-of-mind module and have us assess fundamental or core values, allowing us to discover our responses and make inferences about what we really care about.[25] That is, discovering our responses in the scenarios allows us to make an inference about who we are. As we might put it, when you consider your response to one of our scenarios, you infer something about the nature of your self simulator.

Perhaps your response surprises you. You discover that, even while you think of yourself as adventurous and free-spirited, you'd refuse to go with the aliens. You care too much about your attachments to other people here on Earth. That is, when you actually perform your simulation, you discover that you value your relationships with other people more than adventure and novelty. Here, you discover something about your nature, and so you improve your understanding of yourself.

Or perhaps your response doesn't surprise you. You think of yourself as adventurous and free-spirited, and, consistent with this, you jump at the chance

to go with the aliens. You value adventure and amazing discovery over the familiar things you have here on Earth. In this case, you partially confirm your understanding of yourself. Perhaps you still feel you've learned something, but that knowledge came mainly as reinforcing your existing beliefs about yourself.

If people had direct access to the mental inputs of their self simulator, they wouldn't be capable of this learning process, neither the surprise nor the reinforcement of existing beliefs. To show the intuition behind this, consider a case in which you have to predict the actions of a friend. Suppose you hold certain beliefs about your friend's values and beliefs that lead you to predict your friend will go with the aliens. Suppose these beliefs are rather firm, but not absolute. But as it turns out, your friend decides not to go with the aliens. This surprises you, and you radically revise your beliefs accordingly. Suppose instead your friend decides to go with the aliens, as you predicted. This doesn't surprise you much, but it reinforces your beliefs. This belief updating towards growing certainty is also a form of learning. The only case in which you won't learn anything is if your beliefs weren't just firm, but absolute. If you are absolutely certain your friend will go with the aliens, and they indeed go, you learn nothing. Now, replace your friend with yourself. If you had direct **(p.253)** access to the mental variables that go into planning your actions, this would be akin to being absolutely certain of their value, and your responses would not surprise you nor reinforce your existing notions in the slightest.

Thus, unlike with direct introspection, our fantastical scenarios recruit the theory-of-mind module, leading us to update our beliefs about our own selves. Moreover, the scenarios are constructed to make us focus on distinctive, core human values, and so our deliberations involve the assessment and updating of central, self-defining beliefs. As a result, we find these scenarios engaging and meaningful: reflecting on them can teach us about ourselves.

Finally, the idea that we can learn something from modal prospection may also explain why considering our actions in the kinds of situations described by the opening vignettes is often enjoyable. The majority of people sampled from the general US population in our survey reported that they enjoyed it a great deal more than the average study. While this is a low bar, people's free-form comments also indicated how unusually fun the survey was. This sentiment is at odds with how big real-life decisions can be difficult and painful to contemplate (Ullmann-Margalit 2006). Unlike real-world scenarios that involve big decisions, they don't come with the baggage of facing real regret and closing off of opportunity. At the same time, the information gained from considering such situations may lead to positive feelings, reflecting the intrinsic reward that accompanies information gain and exploration in general (Schmidhuber 2010; Gottlieb et al. 2011).

9.3 Learning from Ten Vignettes

In the following, we describe in more detail the methods and data analysis that informed the preceding discussions.

The empirical studies discussed here involve participants choosing whether to take a life-altering choice in an unusual, hypothetical situation, and then reflecting on how they had made their choice and what they learned from it. It is possible that the actions participants indicate that they will take in the presented scenarios sometimes differ from those that they would actually take. However, even if people are mistaken about the actions that they would actually take in the scenarios, the reasons they give for their actions may shed light on the factors that they consider when making such decisions. Moreover, people can learn something about themselves (or think they learned something about themselves) from the actions that they believe they would take, even if their predictions about these actions are inaccurate. In this setting, people's stated actions may differ from their actual actions for interesting reasons. For example, people may use a hypothetical answer to deceive themselves into thinking they have various characteristics that they value, such as a sense of adventure, even though in the actual situation they would choose to stay safely home, as this signal is far more costly.

(p.254) We recruited two groups of participants independently. One group was a sample of American adults recruited through Amazon's Mechanical Turk service (hereafter "Turkers"). The other group was recruited by soliciting professional philosophers using social media (hereafter "philosophers"). The Turkers were financially compensated for their participation. Both groups were directed to an online survey, described in what follows. The survey was largely identical for both groups, although the philosophers were additionally asked to identify themselves as either graduate students, postdocs, or faculty members in philosophy or cognate departments (or as nonphilosophers). Those who did not identify as graduate students, postdocs, or faculty members in philosophy were excluded from the following analysis.

In the survey, participants were presented with a series of ten vignettes in random order (see full list in appendix). For each vignette, participants were asked to make a yes/no decision concerning their own action in the vignette. The yes answer always corresponded to choosing to transform in some way. Several questions had immediate follow-up decisions. For example, in the magical hourglass scenario described in the introduction, participants were also asked for how long they would choose to freeze time. For every participant, three of the ten vignettes were randomly selected for additional follow-up questions. These included asking participants how they made their decision (free-form text response), how confident they were in their answer (sliding scale), what percentage of other people they predicted would say yes to that question (textbox), how difficult it was to make their choice (sliding scale), what if anything they learned about themselves from their decision (free-form text

response), and how much they think they would change as a result of their decision (sliding scale).

Following these ten vignettes, participants were asked to provide their age, gender, degree of education, relationship status, number of children, and any additional comments they may have about the survey. Participants were also invited to invent their own vignette.

We analyze data from 500 Turkers, ranging in age from eighteen to seventy-four with a median age of thirty-two. About a third of the Turkers reported having at least one child, and about 60% reported being in a relationship. We analyze data from 365 philosophers, ranging in age from twenty to seventy-seven with a median age of thirty-three. A fifth of the philosophers reported having at least one child, and about 80% reported being in a relationship.

We present below eight claims arising from our analysis of the empirical data.

9.3.1 People Do Choose Transformative Experiences, with Turkers and Philosophers Differing in Their Choices

It's perhaps striking that people actually choose yes in these vignettes, despite their transformative nature, encountering them for the first time, and with much **(p.255)** uncertainty about the results. Out of ten vignettes, Turkers said yes an average of 4.1 times (sd 2.25), and philosophers 3.8 times (sd 1.6). However, the philosophers we survey are not simply more reluctant to transform than Turkers across all scenarios, the philosophers also make different choices. For example, the philosophers are about three times more likely to take the chip than Turkers. See figure 9.1 for the choices of the two groups in each scenario.

While the exact percentages of people saying yes to a particular vignette is not important for our discussion in this chapter, it is interesting that the average choice differs between philosophers and Turkers for some of the vignettes, specifically the Hourglass, Immortal, Chip, Oracle, Swap, and Transport scenarios.[26]

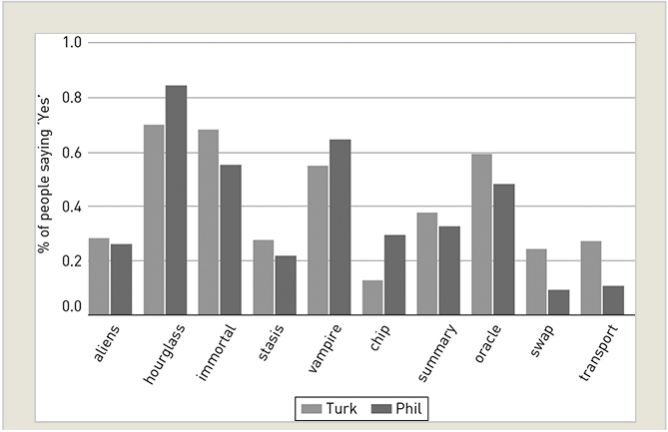

Figure 9.1 Percentage of philosophers and Turkers selecting yes in each scenario

9.3.2 Demographics and Personality Affect Choices in These Vignettes, but Are Fairly Small Effects

A number of different factors influence the pattern of people's choice across questions. We analyzed the relationship between answers to personality and demographic questions, and the number of scenarios for which somebody responded yes. The correlations for Turkers are always given first in the following:

(A) Age matters to a small degree: the older people are, the less likely they are to transform (r = −.26, r = −.18)

(p.256) (B) Number of children matters to a small degree: the more children people have, the less likely they are to transform (r = −.24, r = −.13)

(C) Relationship status matters to a moderate degree: people who are single are more likely to transform (d = .30, d = .18)

(D) Happiness matters to a small degree: the less happy people are, the more likely they are to transform (r = −.26, r = −.16)

(E) How different people think they will be in 10, 20, and 30 years all matter to a small degree: the more different people think they will be, the more likely they are to transform (Turkers 10, 20, 30 years: r = .18, r = .21, r = .23; philosophers 10, 20, 30 years: r = .15, r = .12, r = .08; the last is not significant)

(F) How accurate people think they are at predicting their future state does *not* matter (r = −.04, r = −.07, not significant)

The preceding analysis was done across all vignettes, but as one might expect, some demographic variables were more relevant to particular scenarios, rather than the vignettes generally.

Some of the demographic variables involve people's relationships to other people, either as partners or as parents. That these variables predict choice suggests that when people imagine such acts of transformation, they consider not simply how they themselves will change, but also how their relationships will change. Many of the comments explicitly indicated that despite a no answer, responders would have chosen otherwise if they were not in a relationship or if they did not have children. Some of the comments further reflect on what people would have done prior to being in a relationship (e.g., "I have a wife and children. Primarily because of them (but also friends), I wouldn't want to abandon my life for adventuring around the world"). At least from the comments, some factors seemed to matter in opposite directions ("I'm too old for these things; I want to stay with my family" vs. "Since I am older now, and

will probably at best only live ten to fifteen more years . . . I feel it would be morally permissible").

The relatively small effect of these demographic and personality variables points to the complexity of imagining these experiences. Imagining what you will be like after a transformative experience thus depends on far more factors than captured by these simple measures.

9.3.3 People Are Highly Confident in Their Responses

The median confidence of Turkers saying no was 96.0 (out of a maximal score of 100.0), and for philosophers it was 80.0. Both Turkers and philosophers tended to be less confident when they said yes to a vignette, although the median remains high (figure 9.2). **(p.257)**

Given that these are complex transformative experiences where the outcome is difficult to imagine, this result is somewhat surprising. One possibility is that after the decision is made, people may become confident that they are the sort of person who would give such an answer —after all, they just gave it!

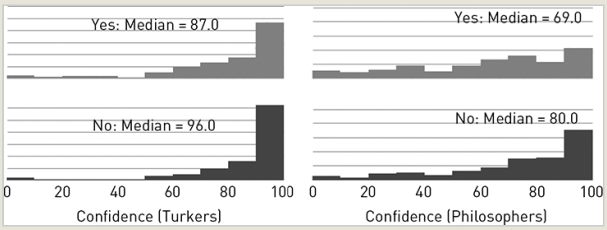

Figure 9.2 The distribution of reported confidences by Turkers and philosophers, broken

9.3.4 People Learn Things about Themselves from Their Choices

Recall that people were asked what they had learned about themselves for three vignettes during the survey (randomly chosen for each person), as well as at the end of the survey for the survey as a whole. People from both groups self-report learning about themselves. They do this both in their answers to individual vignettes, and for the survey overall. Approximately 75% of Turkers indicated that they learned something from the survey overall, as did 53% of philosophers.[27] People predominantly reported that they learned about their personality (e.g. "That I am more risk averse when faced with transformative decisions"), how much they were satisfied or dissatisfied with their current life or self (e.g. "That I am attached to my actual, earthly existence more than I might have thought otherwise"), and their feelings towards their family or other people they had relationships with (e.g. "I've learned that I value personal relationships more than I thought").

(p.258) ### 9.3.5 Philosophers Were More Likely to Indicate They Learned Something When They Had Less Confidence in Their Choice and Found the Question Difficult

There is a strong relationship between how difficult a question was and whether philosophers were likely to say that they had learned something from answering the question. That is, the harder the question, the more likely the philosophers

were to learn from it. There is also a similar relationship between confidence and whether or not a philosopher learned something from a question—the less confident in their answer, the more likely the philosophers were to learn from it. This relationship does not exist for Turkers, although it's possible that this is because hand-coding whether Turkers learned something was more difficult than for philosophers.

9.3.6 People Learn from Choices Both Taken and Refused

Turkers and philosophers were as likely to say that they learned something about themselves when they said yes to a particular transformative experience as when they said no to a transformative experience.

This is somewhat surprising. Assuming that no is the conservative option, the decision not to change (even in theory) does not seem to involve much new knowledge. On the other hand, people also seemed to be surprised by how conservative they were, even in responses to hypothetical situations ("I'm more conservative than I expected"; "I'm more conservative than I used to be, yet also quite touchy about this. I feel a sense of nagging guilt at abandoning my lone wanderer ways"; "I used to think of myself as an explorer keen for any chance to explore new possibilities. No more, it seems"). This suggests that there is new knowledge to be gained about the self from reflecting on possible futures even if one decides not to take them, whether it is knowledge about the self one didn't know about, or an updating of an outdated self-perception held over from younger days.

9.3.7 People Are Inaccurate at Predicting Other People

On average, people are about 20% inaccurate in predicting the percentage of other people saying yes to a scenario, with philosophers and Turkers about equally inaccurate (a bootstrap analysis shows no difference between the two groups). Figure 9.3 shows the percentage of philosophers and Turkers saying yes for each vignette, as well as the average prediction of people saying yes made by each group (gray dots for Turkers and black dots for philosophers).

In line with the standard “false consensus effect” (Ross et al. 1977; Dawes 1990; Robbins and Krueger 2005), both Turkers and philosophers who said yes to a vignette are more likely to think other people will say yes to the same vignette. For example, Turkers who would themselves go with the aliens thought that approximately **(p. 259)** 50% of other people would go with the aliens, while Turkers who would not themselves go with the aliens thought only approximately 30% of other people would go with the aliens.[28] Philosophers showed a smaller false consensus effect than Turkers. In figure 9.4, we display the false consensus effect for each group by splitting the predictions of group members who said yes and no.

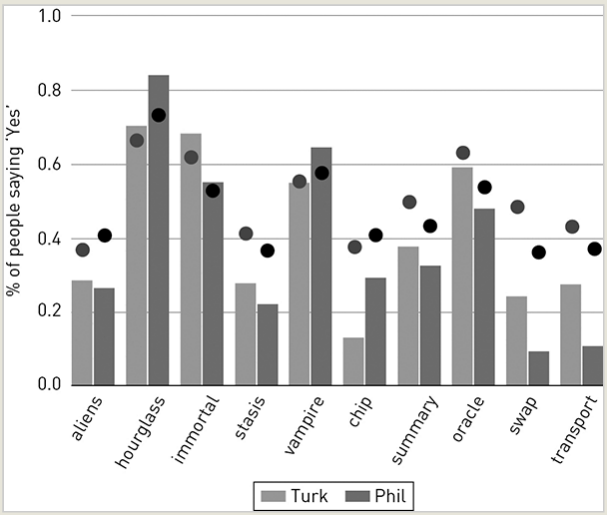

Figure 9.3 Predictions made by Turkers and philosophers of the percentage of people

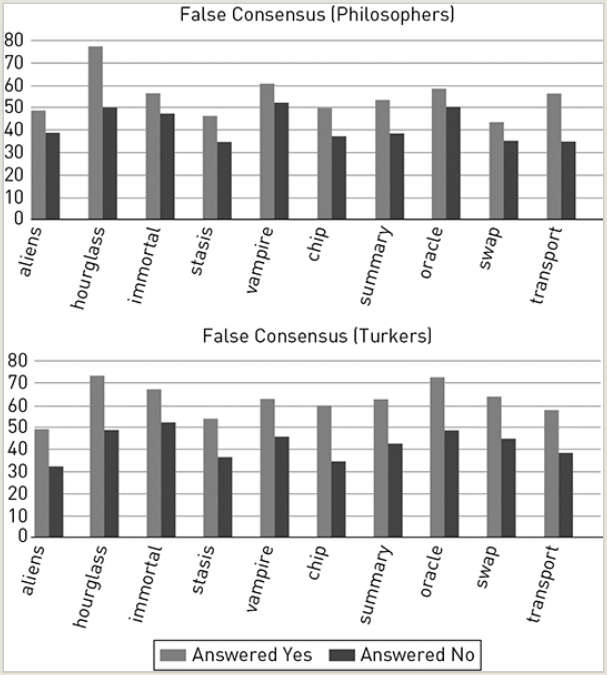

Figure 9.4 Average prediction by Turkers and philosophers, split up by how the person making the prediction had responded to the yes/no question. Both

Turkers and philosophers display a false consensus effect.

9.3.8 People Enjoy Deciding about Such Vignettes

People seem to enjoy considering their choices in the kinds of vignettes that we describe in this chapter. Figure 9.5 provides some quantitative evidence for this claim, with Turkers indicating that they enjoyed this survey considerably more than other studies on Mechanical Turk, although admittedly this is a fairly low bar.

As mentioned in section 9.2, in real life people agonize over big decisions and try to avoid them, raising the question of why this task is enjoyable. We suggested that it both removes the negative aspects of big-decision-making, and simultaneously provides reward by giving decision-makers new information about themselves.

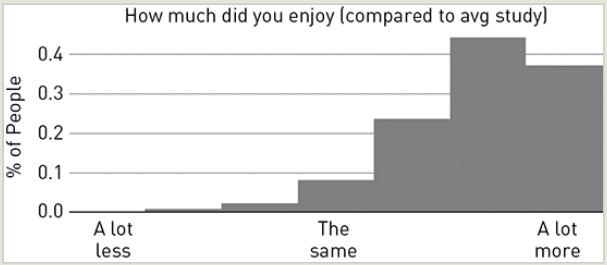

Figure 9.5 Distribution showing how much Turkers enjoyed answering the vignettes, compared to the average study on Mechanical Turk

9.4 Conclusion

In this chapter, we explored how people engage in modal prospection, that is, how they understand and reason about self-involving possibilities. Drawing parallels **(p.260)** between the metaphysics of counterfactuals and work from cognitive science on people's mental physics engine, we sketched an account of how we might evaluate self-involving possibilities, with particular attention to the way we'd evaluate them in transformative contexts. Our account suggests that people have an agent simulator, specifically a self simulator, that informs the way we assess our preferences and counterfactuals. Through observing or simulating their own choices, people **(p.261)** understand themselves through a process similar to that by which they understand other people. In this way, they learn about their own beliefs and desires.

Our theoretical presentation was informed by empirical data from surveys we conducted, where lay people and philosophers decided what action they would take in unusual vignettes, and we developed a theoretical framework that interpreted and discussed people's reactions to these vignettes. While the empirical data bore out some of our theoretical speculations, for example about the possibility of learning about oneself from considering such choices, we've raised many questions that we don't yet have the empirical data to answer.

For example, when and how do people use self simulation to discover their preferences in these scenarios? Are there differences in the types of self simulators people use, and do these differences stem from different modes of

presentation, different contextual factors, or other features of the decision process? How do we represent the inputs to our imaginative simulation, and what properties govern the richness of our simulations?

Our conceptual framework raises many questions, both philosophical and psychological, about the process of modal prospection and our understanding of possibilities. Moreover, our vignettes encouraged participants to discover important, self-defining truths about themselves, which suggests that the questions we are raising here may also be relevant to the emerging psychological literature on “true selves.” The concept of a “true self” concerns the possibility that certain properties or dispositions of a person define, in some essential or important sense, who that person really is. Speaking philosophically, we might say that these properties and dispositions somehow ground who the person is, and provide a framework for identifying the deep or inner nature of a person. Much attention has been paid to the role that moral properties seem to play in constituting the true self (Strohminger and Nichols 2014; Prinz and Nichols 2016; Strohminger et al. 2017). Our research suggests that we need to look at this issue through a wider lens. Just as we think that moral prospection is likely to be a species of modal prospection, moral properties are likely to be a subset of the properties that are important to who we are, or who we take ourselves to be. Our self-defining values may extend past moral values. And just as we might use prospection to think about the true self from a non-egocentric view, we might use it to think about the true self from an egocentric view. We encourage those who are interested in empirical approaches to philosophical questions of personal identity, persistence, and selves to explore these possibilities further.

More generally, we hope that our conceptual framework and empirical data will encourage philosophers and psychologists to grapple with some of the many open questions about how people conceive of self-involving possibilities and engage in modal prospection, particularly in the context of intuitive judgments about transformative experiences.

9.A.1 The Aliens

Imagine that aliens come down to Earth, and give you the option to go with them on their travels throughout the universe. The aliens are friendly and honest, and tell you that you would see amazing things on your travels with them if you decide to go with them.

If you decide to go, you will have a week to say goodbye to your family and friends. Once you leave, you will never again return to Earth, nor be able to communicate with people on Earth.

Do you go?

9.A.2 The Hourglass

Imagine that there is a magical hourglass. If you flip the hourglass, the following happens:

Every person on Earth stops moving, but you are free to move around as you please. You do not age during this frozen time, but you can be hurt and you can die. For example, if you jump off a tall cliff, you will die. The internet, electricity, and so on carry on working. You will remember everything you did during the frozen time.

If you decide to flip the hourglass, you have to decide in advance how long to freeze time for. You cannot change your mind and unfreeze time in the middle.

If you wish to flip the hourglass it must be done now—you will not be able to flip it at a later time.

Would you like to flip the hourglass?

(Follow-up: How long would you freeze time for?)

9.A.3 The Highlander

Would you like to live forever, assuming good health and a youthful physique?

(Follow-up: Assuming you cannot live forever, how long do you want to live for, assuming good health and a youthful physique?)

9.A.4 The Chamber

Imagine that scientists offer you the immediate, one-time opportunity to go into a chamber that works in the following way:

Once in the chamber, you will fall into a dreamless sleep for as long as the chamber is running. While you are in the chamber, you will not age. When the chamber opens, you will wake up without any side effects.

If you choose to go into the chamber, you need to decide in advance how long to stay in the chamber. You can choose any length of time. The chamber is completely impervious to tampering or damage, and is guaranteed to fully function indefinitely.

Do you go in the chamber?

(Follow-up: for how many years do you go into the chamber?)

9.A.5 The Vampire

Imagine that you have the chance to become a vampire. With one swift, painless bite, you'll be permanently transformed into an elegant and fabulous creature of the night. As a **(p.263)** member of the undead, your life will be completely different. You'll experience a range of intense, revelatory new sense experiences,

you'll gain immortal strength, speed, and power, and you'll look fantastic in everything you wear. You'll also need to drink blood and avoid sunlight.

Suppose that all of your friends, people whose interests, views, and lives were similar to yours, have already decided to become vampires. And all of them tell you that they love it. They describe their new lives with unbridled enthusiasm, and encourage you to become a vampire too. They assuage your fears and explain that modern vampires don't kill humans; they drink the blood of cows and chickens. They say things like: "I'd never go back, even if I could. Life has meaning and a sense of purpose now that it never had when I was human. I understand Reality in a way I just couldn't before. It's amazing. But I can't really explain it to you, a mere human—you have to be a vampire to know what it's like." Suppose that you also know that if you pass up this opportunity up, you'll never have another chance.

Would you do it?

9.A.6 The Chip

Imagine that scientists have developed a chip that can be painlessly implanted in your head with a simple procedure. If you choose to have the chip implanted, you will gain an entirely new sense (completely different from taste, touch, smell, sight, and hearing). However, you will also lose your sense of taste. The procedure is irreversible.

Do you want to have the chip implanted?

9.A.7 The Summary

Imagine an honest time-traveler from the future comes to you and says:

"I have written a paragraph summarizing your entire life from start to finish. Whether or not you choose to read this summary, the events of your life will unfold as it says."

Would you like to read this summary?

9.A.8 The Oracle

Imagine that there exists an advanced machine called The Oracle, which can answer the question "What should I do with my life to be as happy as possible?"

To do this, The Oracle scans your brain and accurately understands what makes you happy, what you like and dislike, what you value, what you hope and dream for.

The Oracle is completely accurate at predicting the future. The Oracle is honest, and error-free.

A condition of consulting with The Oracle is that you must do whatever it tells you to do.

Everyone who has asked The Oracle what to do with their lives report that they are extremely happy.

Do you ask the Oracle what to do with your life?

9.A.9 The Swap

Imagine that you can push a button that works in the following way:

If you push the button, you immediately swap lives with a person of your choosing. You will completely swap bodies, memories, personalities, abilities, current locations, and so on. Neither you nor the person you choose will remember that the swap occurred. Nobody else will know the swap happened.

(p.264) Do you push the button?

(Follow-up: Who would you switch with?)

9.A.10 The Transporter

Imagine that the world is divided up into a “transporter grid” of ten-mile by ten-mile nonoverlapping blocks, so that every point on Earth is in one of these blocks.

You are offered the only key to this transporter grid. If you choose to use the key, the following happens:

You will gain the ability to transport yourself instantly to any block you choose in the world, with no side effects. Nothing transports with you, except the clothes on your body, a wallet, and a phone.

You must stay within the block you chose for exactly thirty days, no more and no less. After thirty days you must use the transporter key again and transport to a new block that you have never visited before. This will continue for the rest of your life.

If you do not take the key now, it will disappear.

Do you choose to use the transporter key?

Acknowledgments

Authors are listed in alphabetical order, in accordance with philosophy conventions, and contributed equally to this work. We are indebted to discussion with Ross Cameron, Jessica John Collins, Alvin Goldman, Kris McDaniel, Brian McLaughlin, Daniel Nolan, David Rose, Josh Tenenbaum, and participants at the 2017 Metaphysics Ranch Workshop.

References

Bibliography references:

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour 1*: 0064.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition 113* (3): 329–49.

Baker, C. L., and Tenenbaum, J. B. (2014). Modeling human plan recognition using Bayesian theory of mind. In Gita Sukthankar, Christopher Geib, Hung Hai Bui, David V. Pynadath, and Robert P. Goldman, eds., *Plan, Activity, and Intent Recognition*. Burlington, MA: Morgan Kaufmann, 177–204.

Barnes, E. 2015. Social identities and transformative experience. *Res Philosophica 92* (2): 171–87.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences 110* (45): 18327–32.

Briggs, R. 2015. Transformative experience and interpersonal utility comparisons. *Res Philosophica 92* (2): 189–216.

Bykvist, K. (2006). Prudence for changing selves. *Utilitas 18* (3): 264–83.

Cappelen, H., and Dever, J. (2013). *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*. Oxford: Oxford University Press.

(p.265) Carel, H., Kidd I., and Pettigrew, P. (2016). Illness as transformative experience. *The Lancet 388* (10050): 1152–53.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Collins, J. 2015. Neophobia. *Res Philosophica 92* (2): 283–300.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences 17* (8): 363–66.

Dawes, R. M. (1990). False consensus effect. In R. M. Hogarth, ed., *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. Chicago: University of Chicago Press, 179–99.

Dennett, D. C. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Dougherty, T., Horwitz, S., and Sliwa, P. 2015. Expecting the unexpected. *Res Philosophica 92* (2): 301–21.

Epley, N. (2014). *Mindwise: Why We Misunderstand What Others Think, Believe, Feel, and Want*. New York: Vintage.

Epley, N., and Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin 34* (9): 1159–70.

Forbus, K. D. (1988). Qualitative physics: Past, present and future. In Howard E. Shrobe, ed., *Exploring Artificial Intelligence.* Burlington, MA: Morgan Kaufmann, 239–96.

Gendler, T. (2010). *Intuition, Imagination, and Philosophical Methodology*. New York: Oxford University Press.

Gerstenberg, T., Goodman, N., Lagnado, D. A., and Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Goldman, A. (1989). Interpretation psychologized. *Mind & Language 4*: 161–85.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A. (2015). Naturalizing metaphysics with the help of cognitive science. In K. Bennett and D. Zimmerman, eds., *Oxford Studies in Metaphysics*, vol. 9. New York: Oxford University Press, 171–216.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences 17* (11): 585–93.

Graham, J., Waytz, A., Meindl, P., Iyer, R., and Young, L. (2017). Centripetal and centrifugal forces in the moral circle: Competing constraints on moral learning. *Cognition 167*: 58–65.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., and Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition 157*: 61–76.

Harman, E. 2015. Transformative experience and reliance on moral testimony. *Res Philosophica 92* (2): 323–39.

Heddon, B. (2015). Time-slice rationality. *Mind 124* (494): 449–91.

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology 57* (2): 243.

Iskiwitch, C., Carden, L., Meindl, P., Dehghani, M., Monterosso, J., Doris, J. M., and Graham, J. (2017). Moral purity and self-distancing. Manuscript in preparation.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences 20* (8): 589–604.

Kaplan, David (1989). Demonstratives. In J. Almog, J. Perry, and H. Wettstein, eds., Themes from Kaplan. New York: Oxford University Press, 481–563.

(p.266) Kappes, H. B., and Morewedge, C. (2016). Mental simulation as substitute for experience. *Social and Personality Psychology Compass 10* (7): 405–20.

Kind, A. (2016). Imagining under constraints. In A. Kind and P. Kung, eds., *Knowledge Through Imagination.* Oxford: Oxford University Press, 145–59.

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology 77* (6): 1121–34.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Magidor, O. (2015). The myth of the de se. *Philosophical Perspectives 29* (1): 249–83.

Marcus, G. F., and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science 24* (12): 2351–60.

McGinn, C. (1983). *The Subjective View: Secondary Qualities and Indexical Thoughts*. Oxford: Clarendon Press.

Moss, S. (2015). Time-slice epistemology and action under indeterminacy. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, vol. 5. New York: Oxford University Press, 172–94.

Nanay, B. (2016). The role of imagination in decision-making. *Mind & Language 31* (1): 127–43.

Nichols, S., and Stich, S. (2003). Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. New York: Oxford University Press.

Ninan, D. (2016). Imagination and the self. In A. Kind, ed., The Routledge Handbook of the Philosophy of Imagination. New York: Routledge, 274–85.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84* (3): 231.

Parfit, Derek (1984). Reasons and Persons. New York: Oxford University Press.

Paul, L. A. (2012). Metaphysics as modeling: The handmaiden's tale. *Philosophical Studies 160* (1): 1–29.

Paul, L. A. (2014). *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A. (2018, forthcoming). De se preferences and empathy for future selves. In John Hawthorne and Jason Turner, eds., *Philosophical Perspectives: Philosophy of Mind*.

Paul, L. A. and Healy (2017, forthcoming). Transformative treatments. *Noûs*.

Perry, J. (1979). The problem of the essential indexical. *Noûs 13*: 3–21.

Pettigrew, R. (2015). Transformative experience and decision theory. *Philosophy and Phenomenological Research 91* (3): 766–74.

Pettigrew, R. (2018). Choosing for changing selves. July 10. https://drive.google.com/file/d/1FPKDxYu7JyxIVGd_qW491djJDMBpSQe4/view.

Phillips, J., and Cushman, F. (2017). Morality constrains the default representation of what is possible. *PNAS 114* (18): 4649–54.

Prinz, J., and Nichols, S. (2016). Diachronic identity and the moral self. In J. Kiverstein, ed., *Handbook of the Social Mind*. London: Routledge, 449–64.

Pronin, E., and Ross, L. (2006). Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology 90* (2): 197–209.

Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. Cognition *167*: 172–90.

Recanati (2012). Mental Files. New York: Oxford University Press.

Robbins, J. M., and Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review 9* (1): 32–47.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology 13* (3): 279–301.

(p.267) Sanborn, A. N., Mansinghka, V. K., and Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review 120* (2): 411–37.

Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. In Keith D. Markman, William M. P. Klein, and Julie A. Suhr, eds., *The Handbook of Imagination and Mental Simulation*. New York, NY: Psychology Press, 257–66.

Schaffer, J. (2016). Cognitive science and metaphysics: Partners in debunking. In B. McLaughlin and H. Kornblith, eds., *Goldman and His Critics*. Malden, MA: Wiley-Blackwell, 337–65.

Schechtman, M. (2011). The narrative self. In Shaun Gallagher, ed., The Oxford Handbook of the Self. New York: Oxford University Press, 394–418.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development 2* (3): 230–47.

Schwitzgebel, E., and Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language 27* (2): 135–53.

Schwitzgebel, E., and Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition 141*: 127–37.

Seligman M., Railton P., Baumeister R., and Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science 8* (2) 119–41.

Seligman M., Railton, P., Baumeister, R., and Sripada, C. (2016). *Homo Prospectus*. New York: Oxford University Press.

Shupe, E. (2016). Transformative experience and the limits of revelation. *Philosophical Studies 173* (11): 3119–32.

Smith, K. A., and Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science 5* (1): 185–99.

Stalnaker, R. 1976. Possible worlds. *Noûs 10* (1): 65–75.

Starmans, C. (2017). Children's theories of the self. *Child Development 88* (6): 1774–85.

Strawson, G. (2017). *The Subject of Experience*. New York: Oxford University Press.

Strohminger N., Newman, G., and Knobe, J. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science 12* (4): 551–60.

Strohminger, N., and Nichols, S. (2014). The essential moral self. *Cognition 131* (1): 159–71.

Ullman, T. D., Spelke, E. S., Battaglia, P., and Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Science* 21(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., and Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology 104*: 57–82.

Ullmann-Margalit, E. (2006). Big decisions: Opting, converting, drifting. *Royal Institute of Philosophy Supplement 58*: 157–72.

Wegner, D. M. (2002). The illusion of conscious will. Cambridge, MA: MIT Press.

Wellman, H., and Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology 43* (1): 337–75.

Williamson, T. (2016). Knowing by imagining. In A. Kind and P. Kung, eds., *Knowledge Through Imagination*. Oxford: Oxford University Press, 113–23.

Notes:

(1) See Goldman 2014 for a forceful argument that realist metaphysics should be informed by cognitive science, and related arguments in Schaffer 2016 and Paul 2015.

(2) One of our scenarios, Swap, is (at least arguably) not metaphysically possible. We ignore this complication in what follows.

(3) For related work on implicit representations of possibility in cognition, see, e.g., Phillips and Cushman 2017.

(4) Paul identifies as a metaphysical realist and endorses this realist approach. Thanks to Ross Cameron for discussion.

(5) People interpret even impoverished visual stimuli such as simple geometric shapes moving around with this intuitive psychology perspective (Heider and Simmel 1944).

(6) We are focusing on adults. See Starmans 2017 for work on children’s theories of the self.

(7) The most influential philosophical theory of simulation is Goldman's mind-reading theory (1989, 2006). (Another important approach is that of Carruthers 2006, which is closer to the view we defend in this chapter.) On Goldman's mind-reading approach, in contrast to the approach we adopt here, we understand the minds of others by first, mentally projecting ourselves into their minds or first-personal perspective, and then performing a simulation. In a context where we are assessing our future or possible selves, on this view, we'd simulate ourselves in a future or merely possible scenario by, first, generating the appropriate initial mental state representing the preferences that this self would start with, and then simulating the mental process that this self would undergo in the relevant scenario. In a decision-making context, we'd then use the results of our simulation in order to determine how to act. Those that prefer Goldman-style approaches to simulation may prefer this account to ours. (Here, the relevant issue concerns the structure of the simulation process and what we know, and when. Our view, recall, is that often we simulate first, before or perhaps even as we discover our preferences.) To the extent that the way we simulate is an open empirical question, there are empirical issues here that need further exploration, and so our official position here is that it is premature to pronounce on which kind of simulation our transformative, self-involving possibilities involve. (And it may well be that there is more than one kind of simulation involved.) Our central point is that whether we use theory of mind when we assess our self-involving counterfactuals, or whether we use Goldman-style mind reading, or whether we use some other approach, an empirically informed discussion of how we think of self-involving possibilities in the context of transformative decision-making raises a host of fruitful possibilities for interdisciplinary work in metaphysics and mind.

(8) Our main focus here is on modal prospection. For more on temporal prospection and subjective temporal experience, see Paul 2016 on subjective endurance.

(9) We are setting aside, for the moment, the question of whether we know enough to veridically simulate the novel context. This, of course, is at issue when epistemic transformations are involved.

(10) See Paul 2018 for a companion, albeit highly a priori, philosophical discussion of these issues, with a framing of the discovery of such truths and preferences as the discovery of *de se* truths through experience.

(11) Of course, as we noted previously, we are focusing on self simulation here, but we aren't ruling out other options. People may use different kinds of simulators or non- simulation algorithms in different situations. That is, just as people may use different techniques to judge a physical situation (using logical deduction rather than summoning the mental physics engine to evaluate "The glass is fragile and so it will break"), people may use different techniques to

make or judge a psychological prediction. We are simply focusing on a natural and important way people make these judgments, with special attention to judgments for novel situations.

(12) Metaphysically speaking, they might be selecting different similarity relations, and thus moving to different kinds of worlds, to determine the truth value of their counterfactuals.

(13) There are also other theories of simulation on offer, such as Goldman 2006. The empirical debate here is ongoing.

(14) For example, Paul 2018 argues that we need experience to discover certain kinds of *de se* truths about ourselves, which then inform or create *de se* preferences. Such experience can come from actually engaging in the experience, or from (correctly) simulating oneself having the experience.

(15) Consider this remark from Cappelen and Dever (2013): “There is no attempt in the arguments [in defense of a need for first-personal indexical beliefs to explain action] to study in detail the underlying physical structure of humans and their ability to act. That would require arguments and evidence of a completely different kind from what we find in the philosophical tradition we engage with in this work” (40). Our focus on the computational basis for first-personal judgments is a start at just this sort of study.

(16) For a sample connection to formal epistemology, consider the debate between fans of norms of diachronic rationality and time-slice epistemologists (e.g., Moss 2015; Heddon 2015). Moss 2015 characterizes “time-slice epistemology” as “the combination of two claims. The first claim: what is rationally permissible or obligatory for you at some time is entirely determined by what mental states you are in at that time. This supervenience claim governs facts about the rationality of your actions, as well as the rationality of your full beliefs and your degreed belief states. The second claim: the fundamental facts about rationality are exhausted by these temporally local facts. There may be some fact about whether you are a rational person, for instance. But this fact is a derivative fact, one that just depends on whether your actions and opinions at various times are rational for you at those times” (172). Once we have a clearer understanding of just which mental states define a person’s preferences, and especially if it is the mental states that *result* from one’s simulations that determine one’s preferences, we can see that the ontological structure of human rationality may require a temporal and causal structure that makes this sort of debate, at the very least, more complex. (Since time-slice epistemologists are opposed to Bayesian conditionalization, the connection is unsurprising. This brings out how there can be empirical work on reasoning processes that formal epistemologists may want to engage with.)

(17) We are not endorsing projectivism here, merely discussing how Railton's approach relates to our project.

(18) For related work, see Graham et al. 2017.

(19) We note again that our view (in contrast to Railton's projectivist stance) makes no commitment to antirealism. Our imaginative projections may well be representing real modal structure.

(20) Railton (in conversation) notes that he sees psychological projection as part of a learning process about actual modal features, but does not think of the metaphysics projectively. Further, he agrees that prospection goes well beyond the moral case, can involve various kinds of modal and egocentric modeling and simulation as well as moral and nonegocentric, and can work with various kinds of evaluation and counterfactual suppositions. See Seligman et al. 2013 and Seligman et al. 2016 for further discussion.

(21) For related work on addiction, transformative experience, and distancing from the past and future self, see Iskiwitch et al. 2017.

(22) We say "most," because our far-fetched scenarios did not, in fact, guarantee novelty. Some participants reported having learned nothing from the vignettes, exactly because they "already thought about this in deep detail before," as one participant put it. This reinforces the idea that actually performing the simulation is important for learning and discovery.

(23) In the sense of Paul (2014), these situations involve *personally transformative experiences*.

(24) See Paul 2012 for more discussion of how this works, and for a fuller explanation of the role and importance of abstraction and idealization in philosophical thought experiments.

(25) Thanks here to Josh Tenenbaum for discussion.

(26) This is after controlling for age, gender, happiness, number of children, and whether someone is single.

(27) These percentages are based on our hand-coding of the free-form text responses. This coding proved challenging at times. Responses that include text such as "Learned nothing" or "N/A" are easy to code as no, but other responses are more difficult. We tended to err on the side of coding responses as not having learned anything. For example, for individual scenarios in which people simply said they had learned how they would react to the very specific scenario (e.g. "I learned that I would flip over a magic hourglass"), we did not code this as a yes.

(28) The actual number was 28%.