AFTERWORD: DECISION-MAKING

The central argument of this book is that the epistemically and personally transformative nature of transformative experience creates problems for an individual-level decision procedure based on cognitively modeling or envisioning outcomes of acts involving these experiences.

My argument raises more questions than it answers, and I have been fortunate to receive a wide range of responses to it. In this Afterword, I would like to engage with some of the many excellent comments and criticisms raised by my interlocutors, without assuming that I have preempted or even recognized all of the many dimensions of this issue that one could explore.

Below, I will discuss how problems with transformative experience connect with first personal choosing, counterfactuals, and limitations on empirical projections, informed consent, rational addiction, indeterminate utility, imprecise credences, higher-order computational modeling, and models for unaware agents.

First Personal Choosing

I've been framing big decisions like whether to have a child or whether to become a doctor in very subjective terms, insisting that we think about them as decisions made from the first personal point of view. Although objective moral and social and other factors are relevant to our assessment of decision-making, they have not been the primary focus of my discussion. I set these considerations aside not because objective values are not relevant to any well-considered and thorough deliberation, but because I wish to isolate and critique the distinctively first personal element of our choices.

Big life decisions concerning our subjective futures are most naturally framed, at least in part, as decisions made from the first personal point of view, where we mentally model different possible types of lived experiences for ourselves and then choose between them. These decisions are ordinarily understood as involving, in an ineliminable way, judgments about what our subjective futures will be like if we choose to undergo the experiences involved in the decision to act. When moral and other values are represented first-personally as psychological reasons for making a choice, they can also be represented in the subjective values we assign.

We could argue that we should eliminate the first personal, psychological stance we take when making a choice. But such a reformulation requires us to substantially revise our ordinary way of thinking about such decisions.[1] In effect, we'd have to give up our autonomy in order to preserve our rationality.

As I've made clear, the problem of transformative decision-making is *not* based on the idea that decisions cannot involve considerations drawn from third personal sources such as objective moral considerations, or empirical work in the social and behavioral sciences, when possible choices and outcomes are being evaluated.

Of course we can consult these impersonal sources when assessing alternatives. The worry is not that we cannot or should not do this. Rather, the worry is that, by eliminating the first personal psychological stance, normative constraints on our decision theory would effectively imply that, in contexts of transformative choice, *no* first personal preferences based on "what it's like" considerations can play a central role in the evaluative process. And this is a disaster, for while we are happy to consider moral rules or scientific data when making big and important decisions in our lives, as individuals facing our life's choices, we are not—and should not, for practical as well as

[1] Gilbert (2007), Kahneman (2013), and Wilson (2011) all do an excellent job of showing that revisions are needed in any case.

principled reasons—be happy to simply ignore or eliminate our subjective perspective from its central place in personal decision-making.

There are two kinds of reasons why, when you are making a big decision about your subjective future, you need to make it in a way that retains first personal, psychological assessment.

The first kind of reason is practical. Sometimes there just isn't complete guidance from the impersonal sources: for example, the moral and legal rules may not have much to say about whether you should get the microchip that replaces your sense of taste with a new sensory ability. Or sometimes the morally right thing for you to do depends on first personal considerations, such as considerations concerning what you think it'd be like for your child to be able to hear in a species-typical way. Empirical guidance can also be incomplete. This was brought out in the discussion of the vampire case in chapter 2 by the supposition that the study of vampire psychology is in its infancy. The practical empirical constraint derives from the fact that, when making many actual decisions in the real world, we can't make a big decision about our subjective futures based solely on what experts about psychology, behavior, physiology, and society tell us about how we respond to making certain types of decisions, because we don't have enough of the right kind of information to predict how we'll respond in any suitably detailed way. There are practical limitations on empirical data. (The next section, on the fundamental identification problem, discusses this in more detail.)

A common situation to find yourself in is a decision situation where you know, at least roughly, what the relevant outcomes might be, but you don't know enough about which outcome you'll be mostly likely to experience, or about how you'll respond to experiencing that outcome. You won't know this, because the scientific information you have isn't personalized enough to make accurate individual-level predictions. The science just hasn't developed enough yet. Empirical research, at least in the near future, isn't going to exhaustively *determine* whether you should decide to have retinal surgery or whether you'd prefer being a violinist to being a doctor. This is related to what

is often called the "reference class problem." In low-stakes cases, the reference class problem can be ignored. But in high-stakes cases, cases where your life or way of living is at stake, it prevents you from casually replacing your first personal evaluations with empirical data.

If we had individual-level data that could tell us how likely a particular outcome was for us and how we'd respond to it, then we could argue that big life choices should be made in the same way that we choose not to step in front of a bus or to be eaten by sharks. In cases like the bus or the sharks, we don't need to perform an assessment of the outcome by cognitively modeling what it would be like, because we know what the results would be: we know every outcome is bad, whatever it is like. If empirical work could simply describe the results that were causally determined by our individual-level properties, along with how satisfied we'd be with them, we might want to replace our ordinary deliberative procedure with the expected values specified by these results. In effect, we'd turn the decision over to the experts. But for the sorts of big life choices I've been focusing on, we don't have sufficiently detailed data to do this, and it's not clear we ever will.

The second kind of reason why we need to preserve a role for the subjective perspective in decision-making is principled. When we make big decisions about our futures, we want to deliberate about them This is why normative decision theory matters to us: we are planning out our personal futures and the futures of those who depend upon us, and we want to do it as rationally as possible. But we also want to choose authentically, that is, we want to choose in a way that is true to ourselves, in a way that involves our *self* as a reflective, deliberating person, choosing after assessing our preferences from our first personal point of view and then living with the results. It is natural to think of our point of view and our subjective perspective on the future as an intrinsic part of who we are, where we have control and authority over who we are by making choices that determine what our futures will be like.

Perhaps you disagree. If first personal cognitive modeling creates problems, why not get rid of it? At least in principle, if the data existed,

we could replace first personal cognitive modeling with empirical results and eliminate the role for subjective values and preferences. Or, if the moral rules are known, we can simply use those to determine the decision. Why shouldn't we see such replacement as the truly rational response?

It is important to see how performing this replacement would do great violence to our ordinary way of thinking about deliberation. Again, my argument is not that we cannot consult moral, empirical, and legal sources for guidance, or that we cannot attempt to influence our first personal perspectives using these kinds of sources. Indeed, I hold that when we value revelation and choose experiences in order to see how our preferences will evolve, we should make such choices in concert with our best moral, legal, and empirical standards.

My argument is, instead, that we should not replace our first personal deliberation with what is, in effect, a program that applies an empirically determined or morally determined algorithm to our decisions, so that as agents faced with decisions, we merely feed in an initial possibility and wait for the computer or the scientist (or the philosopher) to tell us how to act.

As individuals facing personal life choices, as real people making decisions about our futures, we don't just want to know what others tell us about the probabilities and values of outcomes, or to perform the computation of the outcome independently of our personal inclinations about the subjective values of the outcomes. We want to know what *we* think and what *we* care about. This is one reason it seems wrong to insist that, to be rational, individuals must make decisions about their future subjective perspectives based solely on third personal considerations such as moral or environmental considerations, at least, when it is given that the lives of others are not at stake and that what is at issue are individuals' own subjective futures. The same point holds when we consider decisions made on the advice of experts. Just as there are many cases in which you should not be rationally obligated to make personal decisions solely on the basis of moral and social and other factors, it is unclear whether you can be

rationally obligated to make personal decisions solely on the basis of the empirical data, replacing your opinion with the expert opinion of others.

If, to be rational, you must determine your subjective values from third personal data alone, you must give up your privileged perspective, the conscious point of view that defines your mental self. If we are forced to give up our first personal view, we're forced to deliberate in a way that implies our preferences determined by our first personal assessment of our possible future experiences don't matter to our rational calculation.

Recall the bizarre idea from chapter 3 that Sally, who has always wanted to have a child, should ignore her personal values and preferences when she chooses. After reading the empirical data and expert (majority) opinion that suggests having a child lowers subjective well-being, she decides to ignore her personal preferences and remain childless.[2] Or consider Sam, who hates the sight of blood, avoids hard work whenever possible and has no interest in medicine, deciding to become a surgeon simply because he knows it is a morally valuable thing for him to do. This way of making decisions seems wrong from our ordinary perspective, because it seems to violate normative rules for deliberating carefully and thoughtfully about how you should choose to live your life.

[2] As Christia Mercer pointed out to me, US society places significant responsibility on the individual who is making the choice. It's not like we're living in a communal society where the responsibilities of childcare are shared: the heavy burdens we lay on the shoulders of parents makes it even more important that individuals think about their preferences for their futures, given that they are the ones who are expected to take on the vast majority of the responsibilities involved. If the response to my argument is that we should be expected to eliminate considerations of our experience when we make big life choices, then the structure of social support needs to change so that responsibility for those choices is no longer an individual matter. Telling someone that they should have a child for reasons of loyalty or society or science and then abandoning them to their future is not an appealing moral or political position.

A further problem with leaving your subjective perspective out of your decisions connects to the Sartrean point that making choices authentically and responsibly requires you to make them from your first personal perspective.[3] A way to put this point is that if we eliminate the first personal perspective from our choice, we give up on authentically owning the decision, because we give up on making the decision for ourselves. We give up our autonomy if we don't take our own reasons, values, and motives into account when we choose. To be forced to give up the first person perspective in order to be rational would mean that we were forced to engage in a form of self-denial in order to be rational agents. We would face a future determined by Big Data or Big Morality rather than by personal deliberation and authentic choice.[4]

Perhaps because we value the privileged perspective we each have on ourselves from our conscious point of view, we want to consider our own subjective preferences and attitudes from our first personal point of view and choose from that perspective (Moran 2001). If we don't choose this way, then in an important sense, we alienate ourselves from our choices, and thus alienate ourselves from our own futures. In other words, if you don't make choices about your future from your own personal point of view, and instead step back and attempt to map out choices based only on some sort of impartial, uncommitted, third personal point of view, you in effect cede authority over yourself.

This holds in two senses. It holds if you disregard your personal point of view entirely. But it also holds if you merely build in your personal point of view as though you were some kind of impartial observer, that is, if you attempt to make a decision by combining third personal empirical information with a third personal stance towards the first personal view of the decision-maker, but where that decision-maker

[3] Moran (2001), p. 164.

[4] Susan Wolf's (1982) "Moral saints," *Journal of Philosophy*, 79 (8): 419–39, brings out related issues.

just happens to be yourself.[5] Neither stance is appropriate for authentic, responsible decision-making, for neither stance incorporates the kind of first personal commitment each of us makes when taking control of our lives and realizing the effects of our decisions.

I must admit that one possible conclusion to draw from my arguments in the main text is that individual decision-makers should, in fact, cede authority to science. After all, science is a guide to truth, and some may be less concerned to preserve first personal deliberation than I. But if this is the conclusion, it needs closer philosophical scrutiny and criticism. If, to be truly rational, we must give up the first personal stance, then meeting the normative rational standard is costly indeed.

The Fundamental Identification Problem

In the previous section, I argued that, for practical reasons, we should not simply replace personal-level psychological assessments of high-stakes outcomes with empirical data. This section explains the methodological underpinning for the practical concern. [6]

Empirical disciplines such as psychology, economics, and sociology collect data on large groups, and subdivide these groups based on externally determined criteria, giving us information about how different types of people tend to respond to different experiences. The research is often focused on broad demographic categories, for example, on what people of different socioeconomic classes, genders, races, and cultures tend to do, often with further subdivisions based on general characteristics like age, health, background, life stage, leisure pursuits, personal abilities, and personality traits. Such information is extremely useful at the impersonal or social level for decision-making, for example, for the development of public or institutional policies, or

[5] Moran emphasizes that we shouldn't take a third personal view on our first personal view. We must *occupy* our first personal view when we deliberate.

[6] I am particularly indebted to Kieran Healy for discussion.

for the assessment of cultural or economic trends. But from the personal point of view, when an individual is making decisions for herself, this information is, at best, only partially useful. This is because the data just do not give us the kind of fine-grained information about how a person who is just like us, a person with just our particular blend of personal abilities and personality traits, our likes and dislikes, our work ethic and neuroses, and so forth, is most likely to respond to a particular experience.[7]

Individuals are complex entities, and each person has a distinctive psychological profile and background, and may respond to an event in a unique way. This means that, even in the same context, different people can have very different personal psychological reactions to having the same type of experience, depending on their individual blend of psychological properties and capacities, even if, when surveying large groups, trends correlated to general characteristics emerge.[8] As a result, when a person is making a decision for herself about her future, she cannot simply substitute social-scientific findings for her own subjective perspective and expect to get a satisfactory result. When it comes to our own choices, the only sufficiently fine-grained guide we actually have to construct our predictions about how we respond to experience is our own self-knowledge and our detailed grasp of our personal point of view.

Of course we often combine our self-knowledge with coarse-grained empirical information to inform and give us guidance, perhaps using it as a rule of thumb when making decisions. We can use the two together, but we cannot *replace* our fine-grained, subjective perspective with predictions from empirical data without far more fine-graining

[7] And complicating all of this, of course, are possibilities like the one we explored in the vampire case, where an individual's preferences are changed by her undergoing the intervention (recall the vegetarian neighbor who loves being a vampire now that she's become one).

[8] I am setting aside physiological reactions, which, when the relevant data exist, admit of more precise prediction and assessment.

than contemporary social science gives us. When considering whether to undergo an experience of a particular type, unless you know just what type of person you are, and which type you fall under that is relevant to the context of concern, and, moreover, that all or virtually all people of this type respond in the same way to having the sort of experience you are considering, empirical results from the social sciences are not sufficient to guide you perfectly to the values for the experiential outcomes you should expect. The problem, in essence, is one of self-location: we don't know enough about the fine-grained details to know which group of respondents we will end up in.[9] The implicit recognition of this fact may be a reason that informed consent is given such a central role in public policies designed for individuals (for more, see the section on informed consent below).

The insufficiently fine-grained nature of empirical information is a version of a well-known problem for social scientists, the *fundamental identification problem*. The fundamental identification problem arises from the fact that the ideal measure of how a given event at time *t* will affect an individual would involve a comparison of the actual individual after she is affected, at time *t+1*, to the counterfactual, unaffected version of herself at time *t+1*:

> The fundamental problem in program evaluation is that it is impossible to directly observe outcome gains for the same person in both states. To deal with this issue, some classical approaches focus only on identifying mean trreatment parameters and rely on conditional independence assumptions to solve the selection problem. Matching has become very popular in this field[10]

The issue here is that the same person can never be observed in different "treatment states," that is, we cannot observe both the actual outcome and the counterfactual outcome of an event acting on a particular individual. Obviously, the problem is that counterfactual

[9] Again, it's related to the reference class problem.

[10] Heckman, Lopes, and Piatek, (2014), "Treatment Effects: A Bayesian Perspective," *Econometric Reviews* 33: 36–67.

outcomes are hard to measure empirically: we can't just go to unactualized possible worlds and gather some data.

In other words, social scientists certainly collect individual level data in the sense that characteristics are observed at the level of individuals. But the fundamental identification problem makes it clear that it is not possible to collect data in the *truly* individual sense that we want when, as individuals, we face our big life choices. The best social scientists can do is evaluate individuals as members of groups with certain identifying characteristics, and compare members of the group who undergo the event to members of the group who do not.

There are various sophisticated ways that contemporary social scientists approach the problem of finding and comparing relevantly matching individuals. The general idea is to try and find the best actual world counterparts for the target group. For example, if the research question involves the effect of living in a poor neighborhood, sociologists may compare an individual living in a poor neighborhood to an individual living in a wealthier neighborhood, where these individuals are as similar as possible with regard to as many characteristics as possible, as long as these characteristics are independent of properties one has as the result of living in a neighborhood with a particular economic status. The idea is that, if the individuals are as similar as possible in the relevant sense, then by pairing them and observing their different trajectories over time, the effect of living in a poor neighborhood is measurable by contrasting it to differences between the properties of the individuals over time. In effect, the premise is that the similarities between the individuals will minimize the possibility of error due to undetected but relevant differences between the individuals. However, while matching similar individuals is an effective research tool, and may be the best strategy available, it and other quasi-experimental methods obviously cannot solve the fundamental identification problem in a principled way.[11]

[11] As Jeremy Freese has pointedly observed: if the matched individuals are so similar when they start out, then why is one person living in a poor neighborhood while the other is not?

The fundamental identification problem derives from the physical fact that empirical information for decision-making in types of actual-world situations can only be developed using comparisons of *different* individuals, with different physical and psychological profiles. It brings out why empirical science must rely on assessments at the level of groups rather than at the level of any particular individual, and why projections for individuals must be constructed from this group-level data. Such projections are based on detectable, measurable, third personal features of individuals who are members of the groups, and sophisticated measures are developed to correct, as far as possible, for relevant differences between individuals within the groups.

The methods used by social scientists are powerful and increasingly sensitive to basic problems of causal inference. They are so successful, in fact, that when it comes to social policy it can be tempting to elide the technical worries raised by the semantics for counterfactuals, the reference class problem, and the demands of the normative standard for rationality. But the perspective taken in the empirical literature is very different from the subjective perspective of the individual, who in personal situations, when making a big decision for herself, has to try to judge, from her first personal psychological perspective, what *her own* responses would be.

As the problems raised by transformative experience bring out, these perspectives can be in tension when policy knowledge meets individual choice at the level of personal decision-making. An individual who faces a life-defining transformative choice is interested in how *she* could be affected by a particular event, not how someone who is merely similar to her is actually affected by that event. And so, since she cares about *her own* actual and counterfactual responses, she measures them, as best she can, by projecting forward into her psychological future and modeling her possible responses to the event in question. This is the deep methodological basis for the practical limits on an individual's replacing her personal perspective with the empirical perspective. Such a replacement will always bring the risk that the social-scientific data she is relying on is too coarse-grained for a decision made at the level of the individual.

In this way, when real individuals in real situations face transformative choices, the abstract and seemingly practically irrelevant questions raised by the fundamental identification problem shift their ground and come into focus as related to central, practical, and life-defining concerns that each of us must tackle in life-defining contexts.

Informed Consent

Imagine facing the following choice.[12] You've been healthy and physically active all your life, with no disabilities or chronic pain. However, just as you retire and start considering Florida real estate, you discover that you need an operation to save your life. Your surgeon tells you that if the operation is unsuccessful you will die, but if the operation is successful you can expect to have a normal lifespan.

You decide to have the operation, but there are further details you need to take into account. In particular, you have to choose between having the operation using a new technique or having the operation using the standard, tried-and-true technique. You know all the available medical facts: your surgeon is the best available, has performed the new technique with a 30 percent success rate, and if the operation is successful you will be pain-free and without any permanent disability. A 30 percent success rate isn't great though. The trouble is, even though your surgeon has a 97 percent success rate using the old technique, a successful operation using that technique will still leave you with chronic, severe pain and disability. You have only one chance to have the operation, because it involves surgery that will permanently damage the organ involved.

How do you decide? One option is to allow your surgeon to choose for you. But since it is your life and death that hang in the balance, you feel that you should make the decision, now that the surgeon has given you all of the information that medical expertise can supply. After all, how he would tolerate chronic pain might be very different from how you'd tolerate chronic pain.

[12] I am indebted to Kieran Healy for discussion.

Your first task is to value the outcomes, and your outstanding problem is to assign subjective value to the outcome of having permanent, chronic pain and disability.[13] The obvious problem is that what crucially matters is how you think you'll respond to living that way, that is, what it would be like for you to live with this pain and disability. For if, for you, living with chronic pain and disability is as bad as, or even worse than, death, then you should choose to have the operation with the new technique. But if living with permanent pain and disability is better then dying, then depending on the details of how much it is for you to live that way, you might want to choose to have the operation with the old technique.

What is knowable in this situation, and what isn't? How, as an informed patient, are you supposed to make this decision? How are you supposed to decide on which technique your surgeon should use?

The question embeds the problem of transformative choice into a context of informed consent. What do we take ourselves to be doing with such consent? The basic idea is that you, as the patient, are supposed to evaluate the different outcomes, determine their values along with the level of risk you can tolerate as a result, and control the decision-making on that basis. But, crucially, then, you need to know *how you think you'll respond* to different possible experiential outcomes.[14]

[13] For simplicity, I am assuming that complicating factors such as wanting to stay alive so as to keep your pension paycheck coming, or allowing yourself to die so that your spouse receives a hefty insurance check, are not present.

[14] Data about how others who have had the operation respond to living with chronic pain and disability won't always help the patient who is attempting to decide whether to have the operation, for patients attempting to make an informed choice also face a version of the reference class problem. That is, how do they know which similarity matters, given all the different classes of respondents that they are similar to in different ways? Which of the classes that they belong to is the *relevant* class that they should locate themselves in? In order to use empirical results to guide your decision at the individual level, you have to know which group of subjects you are most similar to in the relevant respects, that is, you have to be able to locate yourself in the right reference class. But, given the radical changes in experience that result from having the operation, how can you

And in this context, you are unable to get the information you'd need. The problem, of course, is that there is good reason to think that you, with your history, cannot assess the subjective value of living the remainder of your life with severe pain and disability, or know how your preferences could evolve given your different possible experiences in the different possible scenarios. If you cannot determine your values and considered preferences for the outcomes, you cannot rationally determine the act with the highest expected value.The point generalizes, of course, and raises the question: what are we doing when we give our informed consent? What is the individual capacity that is being exercised here that makes a patient's informed consent important and worthwhile?

The structure of this case is relevantly similar to the case of a parent deciding whether her child should receive a cochlear implant, or to a case where, say, a blind saxophonist is choosing to have retinal surgery, and even to a case where a person is chooing to have a child. If, in such cases, we cannot determine the expected value of the act, why think that an informed consent is doing what it is "supposed" to do, that is, why think that the agent is really being given the opportunity to make a choice by being informed in the relevant way? In some cases, such as those involving medical procedures, giving up autonomy and letting the experts decide might be the only thing one can do. But when making a choice of, say, whether to have a child, whether to terminate a pregnancy, or whether to choose cochlear implant surgery for your infant, giving up autonomy is an untenable option. If the justification of informed consent is rooted in the person's ability to understand her values and preferences concerning different possible outcomes, transformative choices pose a challenge.[15]

know which reference class you belong in unless you already know how you will respond to being in chronic pain and being disabled? As Alan Hajek has argued, this problem is just a version of the problem of discovering the right prior probability distribution over the relevant hypotheses (Hajek (2007) "The reference class problem is your problem too," *Synthese* 156: 185–215).

[15] For an interesting and controversial discussion of informed consent, see Neil Levy (2014), "Forced to be free? Increasing patient autonomy by constraining it," *Journal of Medical Ethics*. doi:10.1136/medethics-2011-100207.

As I noted, there may be some reason for a patient to want to voluntarily give up autonomy when faced with medical decisions involving uncertain outcomes. There are also related questions about how to understand patient decision-making and autonomy in the sociological literature. As work on the exchange of human blood and organs shows, there is a way of understanding the role of informed consent as a practice designed to give patients the *impression* of rational agency and control over the outcomes, a practice that can even involve encouraging patients to visualize or imagine fictional outcomes of their decisions in order to suggest a way of subjectively valuing them. The reason for encouraging patients to envision fictitious outcomes is that patients who make their choices under the impression that they have a sense of what the outcomes will be like, even if the sense is grounded in purely fictitious "facts," may in fact respond better to the actual outcomes, however they turn out.[16] In other words, an implicit recognition of our epistemic inability to evaluate transformative experiences may be subtly embedded in certain deeply practical strategies involved in managing the psychological framing of patient decisions and donor exchange.

Rational Addiction

My reflections relate to discussions of rational addiction,[17] along with ideas about "pre-commitment devices" and future discounting.[18] If a potential addict can't know what it will be like to have the high that

[16] Kieran Healy (2006) *Last Best Gifts*. Healy's work in this area focuses on the procurement strategies involved in human blood and organ exchange. He notes that "this might suggest that organizations manage donors as ants tend to aphids, farming them assiduously and stroking them for nectar as needed. But individual expectations and organizational capacities have coevolved" (p. 113).

[17] I'm indebted to John Quiggin for discussion.

[18] See Gary S. Becker (1998), *Accounting for Tastes*, and Jon Elster (1979), *Ulysses and the Sirens: Studies in Rationality and Irrationality*, and Elster (1997) "More than enough" for discussion.

could get him addicted, how can he, in principle, rationally choose to become addicted, or rationally choose to avoid addiction?

Standard models of rational addiction assume that the potential addict has fixed and unchanging preferences over his lifetime consumption patterns. He can predict with perfect foresight how his preferences at a given time in the future will change as a result of his choices today. If the potential addict knows that he'd prefer having the preferences of the addict to his current preferences, he can compare his options. But if this feature of the model is removed, in certain contexts, the problem of transformative decision-making arises. In such cases, the key fact is that without the knowledge of what it will be like to enjoy the high so much that you desperately want it to continue, the addict cannot make the comparison between his pre-hit self and his post-hit self.

And, indeed, we can imagine a case involving a drug that affects people in variable ways. Ninety-nine percent of those who try it do not get addicted with the first hit. But 1 percent do. In other words, this drug is so powerful that, for those of you among the unlucky 1 percent, once you experience the high, you have an intense and overwhelming desire for *more*, and you don't care about anything else. All you want is to experience that burning, bright, intensely beautiful high.

Now imagine someone who has never used this drug, but who really wants to try it. He just wants to have one hit to see what it's like. (He wants to know what all the fuss is about.) He has never experienced the kind of intense high that this drug provides—indeed, that's part of the attraction. Perhaps he's had experience with a wide range of other drugs, and never got addicted or even felt the need to try them more than once. So he thinks that he can handle a bit of experimentation, that is, given his previous experience, he thinks he won't get addicted with the first hit.

The trouble is, that until he experiences the particular, gripping, and distinctive kind of high that this drug gives, he cannot know how intensely it will affect him. He has to have the experience of the high in order to be able to judge whether it is so intense that he is the kind

of person who will become addicted. But, once he knows how it will affect him, it will be too late. The problem is not just one of uncertainty, it's one that involves the agent being unable to grasp the relevant content. Like ordinary Mary in her black-and-white room, he is not able to know what the experience is like, but since what the high is like will determine his future preferences, it is not possible for him to determine these future preferences without having the experience.

As it turns out, once he takes his first hit, he experiences a pure and intense longing for another hit, a longing like nothing else he's ever experienced. While the high lasts, he cares for nothing more but continuing the high itself, and as it fades, he is seized with the desire for more, and will do anything to get it. If his new preference structure remains, that is, if he is in fact addicted, then he is the victim of a transformative choice: he acted rationally given his pre-addiction preferences to have only one hit (just to see what it was like), but the high changed his preferences in a way that he could not foresee.

We don't have to try very hard to imagine such a drug: the popular perception of crack cocaine is of a drug that is so powerful that some people can get addicted the first time they use it.

Indeterminate Values

Recent work by Alan Hájek and Harris Nover (2004, 2006), develops a case based on a gamble they call the *Pasadena game*, and argues that this is a case where we cannot assign an expected value to playing the game. In the Pasadena game, because of the way the game itself is structured, the expected utility, that is, the expected value, is indeterminate, leading Hájek and Nover to conclude that a value cannot be assigned to playing the game.[19]

The result is related to well-known decision-theoretic puzzles with cases like the *St Petersburg game* in which the expected value of playing the game is infinite. In that game, a fair coin is tossed until it comes up

[19] I am indebted to Alan Hájek, Andy Egan, and Kenny Easwaran for discussion.

heads, and you receive exponentially escalating payoffs. The longer it takes to get heads, the higher the payoff, and since the increase in payoff is exponential, the payoff increases rapidly. To calculate the expected value of the game, you sum the probabilistically weighted values for each of the possible outcomes and get the mathematical result that the expected value of playing the game is infinitely high.

The problem the St. Petersburg game raises for expected utility theory is that, if the expected value of playing the game is infinite, standard theory seems to tell us that you should be willing to pay everything you've got (assuming your worth is finite) just to play the game once. This seems wrong, and raises questions about the ability of standard decision theory to handle cases involving infinitely valued outcomes, at least when it is thought of as providing the normatively rational standard for making decisions where agents start with assigned values and probabilities and then calculate expected utility results.

The Pasadena game is different from the St. Petersburg game but raises a related problem. In the Pasadena game, according to Hájek and Nover, the expected value of playing the game is not infinite; rather, the value is mathematically indeterminate. In this game, a coin is tossed until it comes up heads, the possible payoffs grow without bound, and the payoff for the first time the coin comes up heads alternates: positive values (rewards) alternate with negative values (costs). The indeterminacy comes from the peculiar structure of the Pasadena game, for, as it turns out, the expected value of playing it depends on the ordering of the payoff table: depending on how the alternating terms of the payoff are rearranged, the expected value of any particular play can be infinite, or if the terms are rearranged, it can diverge all the way to negative infinity. In more technical terms, the expectation series of the game "conditionally converges." Hájek and Nover conclude that the value of playing the game is mathematically indeterminate, and the upshot, on their view "is that there is no fact of the matter of how good the Pasadena game is, at least in terms of expected utility, and hence in terms of standard decision theory" (Hájek and Nover 2006, p. 705).

Hájek and Nover conclude that the Pasadena game and its variants pose a deep threat to standard decision theory, where standard decision theory is understood as an approach to rational decision-making where agents start with values (or utilities) and probabilities in order to evaluate choices.

Now, Hájek and Nover's conclusion is controversial, since it depends on accepting that the game itself is coherent. Critics have suggested alternative approaches.[20]

But whether or not the game is actually coherent, I draw the lesson that *if* the game is coherent, and *if* its expected value is indeterminate or unassignable, then standard decision theory faces a problem. If there exist coherent games with indeterminate expected values, either standard decision theory cannot handle such games and needs to be revised in order to handle them, or there is nothing to be said about how good such games are.[21]

[20] Fine (2008) "Evaluating the Pasadena, Altadena, and St. Petersburg gambles," *Mind*, 117 (467): 613–32; Colyvan (2006) "2006: No expectations," *Mind*, 115 (July), pp. 695–702. Kenny Easwaran argues that there are intuitive reasons to value the Pasadena game using its weak expectation. This means that there are at least some ways of playing the Pasadena game that give a determinate, very low expected value (less than a dollar!), suggesting that there is hope for standard decision theory in some contexts (Easwaran (2008) "Strong and weak expectations," *Mind* 117, 633–41). Since there is no parallel expectation for cases of epistemic transformation, solving the decision theoretic problem for transformative choices requires some other strategy.

[21] To the extent that decision theory gives us guides for rational action, where such guidelines are modeled by playing various sorts of games, this result raises important questions about the ability of decision theory to handle cases with indeterminate expected utilities. If we are confronted with structure in some decision-theoretic context that is relevantly similar to the structure of such games, it would seem that, without significant modifications, decision theory cannot be used to make a rational decision. That is, there is no rational way to make sense of playing a game with that structure, and so no rational way to maximize expected value in a decision case with that structure.

Why do games matter? Because decision theorists use games to model decision problems and the strategies we should use to rationally approach them. So, the conditional lesson from the Pasadena game should be generalized: if there is a decision problem with indeterminate expected values, either standard decision theory cannot handle such a decision and needs to be revised in order to handle it, or there is nothing to be said about how rational the decision is.

You might think this is all just fun and games. That is, you might think these games are interesting and fun to think about, but that there is no relevant structural similarity to cases we actually care about: these games just concern esoteric mathematic possibilities generated by complex rules and artificial contexts. They aren't something that agents making personal decisions would ever confront.

But, in fact, the nonmathematical cases we've been discussing, like the case of choosing to try durian or choosing to become a vampire, share an important similarity with these games. In particular, cases of decision-making involving epistemically transformative experiences can share the same deeply problematic feature that, in the Pasadena game, gives the undesirable result: the expected values are indeterminate.

Recall the description of subjective decision-making from chapter 2, where you engage in a kind of cognitive modeling from the subjective perspective. To make a choice, you construct a mental model of the situation, thinking of the different options, running a mental simulation of what it would be like, and assigning it a subjective value based on what it seems it would be like. You then assess and compare these values to make your choice.

Transformative experiences throw a wrench into this process, because they are epistemically transformative. If an option involves an epistemically transformative experience, you lack the parameters you need to run the simulation for that option. So if you can't run the simulation, because you can't know what it is like, then you can't assign it a value.

Consider the *durian game*: you are visiting Thailand for the first time, and need to choose between having a piece of ripe pineapple and

having a ripe durian for breakfast. You've never eaten a durian, nor anything resembling it. To play the game, you have to choose and eat a fruit for breakfast. To win, you have to choose the kind of fruit that you'd like the most. So you have to choose between having pineapple and skipping durian, or having durian and skipping pineapple. If you choose pineapple, and it tastes better than the durian would taste, you win. If you choose pineapple, and it tastes worse than the durian would taste, you lose. If you choose the durian, and it tastes better than the pineapple would taste, you win. If you choose durian, and it tastes worse than pineapple would taste, you lose.

It should be obvious that, unless you know what it would be like for you to taste a durian, that you cannot play this game if you want to choose based on what you think the taste will be like. Until you taste a durian, you have no idea whether you'll like durian more or less than pineapple. The upshot is that there is no fact of the matter of how good the durian game is, at least in terms of expected value, and hence in terms of standard decision theory.

Now we can tweak the game a little bit: imagine the value of winning the game is tied to the intensity of your pleasure in eating the fruit, and the cost of losing the game is tied to the intensity of your disgust in eating the fruit. This makes the game more complicated, because, in principle, you can win big or win small, and lose big or lose small. If you choose pineapple, and it tastes much better than a durian would taste, you win big. How big? It depends on how much better, relative to the durian, the pineapple tastes. On the other hand, if you choose the pineapple and it tastes only a little better than the durian would taste, you win small. If you choose pineapple, and it tastes much worse than the durian would have tasted, you lose big. If you choose the durian, and it tastes much better than pineapple would taste, you win big, and so on.

When faced with a durian game whose values are tied to the relative degree of pleasure or disgust you experience with the relevant outcomes, it becomes clear how ineffective decision-theoretic models are for cases like this. For not only a re you unable to judge whether or not you should

play the game, for any outcome that requires a comparison of the taste of durian with something else, you can't judge what its value is, and you can't judge its relative value either. So for any such outcome, if you have never tasted durian, you cannot rationally assign that outcome a value—that is, for any such outcome, its value is epistemically indeterminate.

We find ourselves with the very same problem we had when considering the Pasadena game. No model of decision is available to the game-player who chooses based on what she thinks it will be like to taste durian, for, just as in the Pasadena game, she cannot determine the expected value of playing the game. This is particularly apt when thinking of rational choice in terms of credences, subjective values, and epistemic utility theory. If the agent cannot determine the expected value of playing the game, standard models of epistemic utility don't apply. Or, put another way, they fall silent on the value of the game.

The basis for the similarity between this feature of the Pasadena game and cases of epistemically transformative experience has nothing to do with the mathematics of infinitely valued outcomes. Rather, it's because both cases put the decision-maker in a position of epistemic indeterminacy. The mathematical indeterminacy of the value of the Pasadena game leads directly to epistemic indeterminacy: it entails that there is no way to rationally assign a value to playing the game. In decisions involving epistemic transformative experience, the facts about experience entail that, given the way the decision is structured, the values of the relevant outcomes are epistemically indeterminate. The source of the epistemic indeterminacy comes from the nature of experience-based knowledge rather than mathematical indeterminacy, but it is the fact of epistemic indeterminacy itself that matters, not where it comes from. The decision-theoretic implications are the same.

Consider, now, the *baby game*: you've never had a child. To play the game, you have to choose whether to have a child. To win, you have to choose the act with the results that you'd like the most. So you have to choose between having a child and passing on the joys of a childfree life, or having a childfree life and passing on the joys of being a parent. If you choose to have a child, and being a parent has a higher

subjective value than living the childfree life, you win. If you choose to have a child, and being a parent has a lower subjective value than living the childfree life, you lose. If you choose the childfree life, and being childfree has a higher subjective value than being a parent, you win. If you choose the childfree life, and being childfree has a lower subjective value than being a parent, you lose.

But unless you know what it would be like for you to have a child, you cannot rationally play this game, because you cannot rationally value the outcomes. And here, given the personally transformative nature of the experience, you cannot even model how your values and preferences would change if you play the game one way (having a child) versus playing it another way (remaining childfree). The upshot is that there is no fact of the matter of how good the baby game is, at least in terms of expected subjective value, and hence in terms of standard decision theory.

The consequence of all this is that if we wish to preserve transformative decisions such as whether or not to have a child (or to become a violinist, or to try a new drug, or to get a cochlear implant, and so on) in the way in which I have been framing them, as subjectively based decisions necessarily involving considerations of one's subjective future, either decision theory understood as expected value theory needs to be revised, or it must be limited to the decision problems it can fit—or we must grant that *this is not the right way to play the game*.

Finkish Preferences

The Pasadena game connects with the first part of the problem with transformative choice, the epistemic problem of subjectively valuing outcomes. The second problem of transformative experience, the fact that having the experience changes your personal preferences, connects with well-known puzzles involving dispositions. [22]

[22] I am indebted to Richard Pettigrew, Matt Kotzen, and John Collins for discussion.

If an experience is personally transformative, and you know how it will personally transform you, you can know how to simulate your different possible lived experiences and assign them values in a way that accommodates your future preferences, even before you have the experience that changes you.

But when the experience is, at once, both epistemically and personally transformative, you can't predict what simulations you will need to run, since your parameters change in virtue of having the experience itself, and you can't know how you'll be changed until you have the experience. Having the experience of performing the act and its immediate consequences will change your assessment of the expected value of your act. Unless and until you have the relevant experience, you don't know which simulations you'd have to run (or how you will value them). So unless and until you have that experience, you can't assess its expected value, because your ability to make that assessment depends on your having had the experience.

A way to put this is that you, as a cognitive simulator of possible outcomes, are *finkish*.[23] A finkish simulator is a simulator whose original disposition to simulate disappears when it undergoes the experience: it changes its disposition just when the experience is had. As a finkish simulator, you change how you simulate when you are put to the experience. You have preferences about your possible future as a vampire, and those stay the same, so long as you are not bitten. But when you are bitten, because what and how you value changes, how you'd simulate changes, and thus your preferences change.

The problem here, of course, is that you are supposed to perform a simulation in order to decide whether to have the experience in the first place. But because you are finkish, before you have the experience you should perform one kind of simulation, because of the kind of preferences you have at that time. But when you have the experience,

[23] "A finkishly fragile thing is fragile, sure enough, so long as it is not struck. But if it were struck, it would straight away cease to be fragile, and it would not break." Lewis (1997) "Finkish dispositions."

you change, such that the simulation you should have performed is different. And you only change your preferences about your acts *because* you were put to the experience: it is the experience itself that brings about the change. You have finkish preferences.

Now, add to the finkishness the fact that you can't know *how* you will fink. As I discussed in chapter 2, because you don't know what it will be like to have the experience, you can't employ higher-order techniques to try and simulate the new and different outcomes and their values that would determine your post-experience preferences. You can't, as it were, before you've had the experience, put yourself in the shoes of the post-experience person (you after the experience) and run those outcomes, because you don't know what values that post-experience-you will assign. In other words, in this sort of case, you are transformatively finkish.

So, if the experience is epistemically and personally transformative, you, as the simulator, are in decision-theoretic trouble, because you are a finkish simulator. When you face a transformative choice, then, you face the fact that you are a *transformatively finkish simulator*. If an agent has transformatively finkish preferences, then, she cannot satisfy van Fraassen's (1984) Principle of Reflection.[24]

Imprecise Credences

Instead of reformulating transformative choices in terms of revelation, could we respond by relaxing the normative standard for decision-making? We already have independent reasons to think the normative standard needs changing, because we have reasons to think that cases where we don't have precise credences need special treatment. Perhaps the real problem with transformative choices are that they are members of the class of examples involving imprecise credences.[25]

[24] Bas van Fraassen (1984), "Belief and the will," *Journal of Philosophy* 81 (5): 235–56. According to van Fraassen, rationality requires that your credence in proposition *P* at t1 is your current expectation of your credence in *P* at t2.

[25] I am indebted to Richard Pettigrew for discussion.

Let's look at this using the example of choosing to have a child. Here, you are choosing whether or not to have a child, basing your decision on whether you'd prefer to become a parent or whether you'd prefer to live your life childfree. The suggestion above is to understand this transformative choice as a problem stemming from the agent's inability to assign precise credences. So, for example, the suggested interpretation for understanding the epistemic difficulty you face before you have your first child is that it is due to the fact that you simply don't have enough evidence to warrant any particular assignment of credences to the states needed to bring about the range of outcomes associated with having a child. Put slightly differently, the problem is framed as a problem where, when facing the choice, we find ourselves faced with a range of possible subjective values associated with having a child, and we don't know which credences to assign to the states of the world needed for the outcomes in this range.

Understood this way, the problem about choosing to have a child comes from epistemic ambiguity concerning which act will maximize expected value, since acts are functions from states to outcomes, and you don't know which credences go with which states. If so, various decision-theoretic models have been proposed for this situation. In particular, the right response to the lack of evidence you have about what it will be like to have a child might be to think of your doxastic state as modeled by a *set* of probability assignments to states and their corresponding outcomes, rather than by a single probability measure over the possible states.[26]

[26] There is no standardly accepted approach to this problem, and there are two main proposals that are hotly debated: (i) whether we need imprecise credences at all, and (ii) whether they create problems for decision-making, for example, by making us vulnerable to money pumps. See White (2009), "Evidential symmetry and mushy credences," in Gendler and Hawthorne (eds), 161–86; Elga (2010) "Subjective probabilities should be sharp," *Philosophers' Imprint*, 10 (5): 1–11; Moss "Credal dilemmas"; Carr "Imprecise evidence without imprecise credences," unpublished MS; Joyce (2010) "In defense of imprecise probabilities in decision-making and inference," *Philosophical Perspectives*, 24 (1) :281–323.

In other words, perhaps the problem is that, in one frame of mind, you think that you know the subjective value of what it's like to have a child is very high, say, when you are imagining yourself joyfully cradling a cooing newborn, and you believe an outcome with a high subjective value has a reasonably good chance of obtaining, given your current credences. So you assign credences such that having a child maximizes your expected subjective value. But in the dead of night, you find yourself in a different frame of mind. Perhaps you imagine a constantly screaming, colicky child, and the relentless drudgery of changing diapers, combined with months of sleepless nights, or perhaps you dread outcomes where you have a disabled child, such as a child with Down Syndrome or a child with dwarfism. . . and then assign credences such that remaining childless maximizes your expected subjective value.[27]

And so on, such that you have a set of different ways of assigning credences to states of the world, where these different ways of assigning credences result in different, conflicting ways of maximizing your expected subjective value. Perhaps the problem with determining what it is like to have a child is really this one: we simply don't know which way of assigning credences is the right one, and the evidence we have is not enough to warrant assigning *any* particular assignment of credences. So your set of ways to assign credences must include many, many possible way of assigning them.

If you *could* actually know what the different subjective values were for your different possible outcomes, this would be an interesting way to model the situation you'd be in. Once we could assign subjective values to the outcomes of what it is like to have a child, we'd have the subsequent problem of determining which credences to attach to which states of the world, and it is plausible that this problem would involve issues with imprecise credences. And, in fact, some people who have not had a child do seem to think about the problem this way, that is, they agonize over which way they should assign their credences in order to decide what to do.

[27] Moss, "Credal dilemmas."

If this were our situation when we face the choice of whether to have a child, I still think we'd be in pretty deep and rather interesting trouble, especially because this is such a central, life-defining decision that people want to be able to face rationally—and there is no consensus about how to handle imprecise credences. But this is *not* the situation, at least not in the first instance. For, as I've argued above, you *don't* actually know what the different subjective values will be for your different possible outcomes. How you will value the outcomes is determined by what it is like for you to generate and stand in the attachment relation to your child. Moreover, until you know what it is like to have your child, you can't even know the *range* of the subjective values. That is, you don't know just how wonderful it could be to have your particular child, and all the experiences that follow from this, or just how awful, or heart-wrenching, it could be (if, for example, your child is born severely disabled, in great pain, with only a few months to live). How bad, or how good could it be? Possibly *very* bad, or possibly *very* good––but that's about all you know. You don't even know if the range of outcomes is symmetrical, that is, that whether the best possible outcome is as high on the scale as the worst possible outcome is low. Perhaps the value of the worst possible outcome is very, very bad, while the value of the best possible outcome is merely good, or quite good. Moreover, as I discussed in the main text, because generating and standing in the attachment relation to your child changes your preferences, this calls into question *any* model of your decision that is based only on the preferences of the pre-experience self.

The main problem, then, is the same problem we had in chapter 2, when we discussed the possibility of becoming a vampire, or the possibility of having a chip implanted in your brain that would eliminate your sense of taste while endowing you with a new sensory ability that is stunningly different from the usual five. In the sensory capacity case, you must choose to discover a new kind of sensory ability in place of the ability to taste, or you choose to retain the ability to taste instead of discovering a new kind of sensory ability. The problem here is that you can't compare having the new sense ability to what it is like

for you to have the sense of taste in order to decide which one you'd prefer. You can't even make a decision based on comparing the range of the values of having the new sensory ability to the range of values of tasting things, because without actually knowing what it is like to have the new sensory ability, there isn't any useful sense in which you can know the range of the values of having it. And finally, you face the central problem of transformative choice: that you don't know how having the new sensory experiences could change your preferences.

That is, knowing what it is like to have a new sensory ability by actually deploying the sensory ability in experience is precisely what determines your judgment of how subjectively valuable, from your point of view, it is to have it, and is what would give you the ability to predict how it would change your preferences. By extension, knowing what it is like to have the new sensory ability by actually deploying the sensory ability is necessary for you to know whether you want to trade in your ability to taste. The experience of having a new sense is simply too radically different from your previous experiences, and too potentially life-changing, for you to be able to assess the range of the subjective values of having it and the way your preferences could evolve before you know what it is like.

Like the experience of gaining an entirely new sensory ability, you must experience what it is like to have your child before you can know the range of subjective values involved, how to compare the subjective values, and what your new preferences might be.[28] If you don't know the subjective values of your outcomes, or the range of values, or how your preferences will evolve, then decision-theoretic models for imprecise credences are of no use, for the problem, at least in the first instance, is with knowing the subjective values and how to handle the change in preferences, not with assigning degrees of subjective belief.

[28] Andrew Solomon's *Far from the Tree* gives an excellent account of how the preferences of parents with disabled children can evolve.

But does this mean we are completely stuck here? Are the tools from formal epistemology of no use at all? Let's step back and look at the choice to have a child again, and see where formal epistemology can take us.

Is there anything we know about the situation that we can exploit to make progress on the problem of transformative choice? Before having the transformative experience, we don't know the individual subjective values of the outcomes, we don't know the range of these values, we don't know how to compare them, and we don't know how to predict our preference change. But let's assume we do know, in the choice-to-have-a-child case, that the possible outcomes could be very good, or very bad, or something in between. So even if we don't know exactly what the range is or how to compare the values, we know the values are possibly very positive or very negative and (let's also assume) they are comparable.

What this means is that small changes in information can have big effects, because knowing even a little bit more about a high-value outcome can have a significant effect on calculations of expected value. Even a little bit of information about the values of these outcomes or the credences we should attach to them could result in a significant change in expected subjective value.

For example, if you were able to find out that the range of the values was asymmetric, such that the worst outcomes were very bad, while the best outcomes were merely pretty good, and you were also able to discover how they were comparable, then if you adopted a rule that current preferences were to receive priority, you could exploit a version of the models for decision under ignorance that we discussed in chapter 2. In such a situation, you would be rationally permitted to choose to remain childless, because the asymmetric shape of the value space is so heavily tilted towards the negative.

On the other hand, if you knew that, perhaps because of the psychological changes that having a child wreaks in a parent, that the range of the subjective values was asymmetrically tilted towards the positive, and you adopted a rule that post-experience preferences were to receive priority (and relied on the testimony of satisfied friends and relatives),

perhaps you could be rationally permitted to choose to have a child merely on this basis.[29]

Do we have this information about the shape of the value space? In particular, can we exploit the psychological facts about preference change (setting aside, for the moment, the worry about whether we can rationally choose to change our preferences this way) to know that, whatever it is really like to have a child, we will find intense, revelatory, positive subjective value in it, and so the negative outcomes are minimized? Or, on the other hand, can we know that the worst outcomes are very bad, but the best outcomes are only pretty good?

No. And the reason, again, stems from the basis for the cognitive phenomenology of having a child, which derives in part from the attachment formed between the parent and child. Forming this attachment is part of the ground for the revelatory nature of the experience, and also for the change in preferences that parents experience. A direct consequence of forming this attachment, and the revision of preferences it entails, is a new and deep emotional vulnerability, one which allows for a wide span of values.

Once she exists, you care very much about your child, and about what happens to her. And in particular, given the attachment you form to your child, if she is born severely disabled such that she experiences great pain and suffering, or dies in infancy, you will experience an indescribable amount of emotional and psychological distress. The specter of the suffering and death of one's child engages a parent's greatest fears. Yet, if the horrors associated with this sort of outcome are not realized, and all goes well, parents report feeling an intense joy, rating having a child, at least anecdotally, as one of the most rewarding experiences of their lives.

This strongly suggests that either there are no exploitable asymmetries, or that we don't know enough about the nature and shape

[29] I'm making a lot of big assumptions here about the legitimacy of adopting the decision rules about preferences, relying on testimony, and so on. But it's worth seeing how far we could go under such assumptions, even if I don't think it will pan out in the end.

of the value space to know whether there are such asymmetries, to rationally choose to have a child based on what it would be like. If you had more information about what it would be like for you to have your child, you might be able to determine enough about the value space to find a useful asymmetry and to determine how your preferences would evolve. Once you had this information, if you had even a small amount of information about how you should assign credences to the states needed for the different subjective outcomes, you might be able to use this to determine any significant asymmetry in the expected values of your acts. When the stakes are high, even a small asymmetry in probability can be significant. But in most real-life cases, you don't have what you need to get started.

We might put it this way: the main problem with truly transformative choices is not a problem in formal epistemology; it is a problem in formal phenomenology. To make a start on the problem, we need to think about the possibility of developing models for epistemically indeterminate values, perhaps starting by examining ways to model imprecise values (although since the real problem is the epistemic inaccessibility of the values, models for imprecise values won't be able to do all the work). Imprecise credences, at least in the first instance, are not the problem, although if we solve the formal phenomenology problem we'll still need to address the assignment of imprecise credences and give an account of how to manage temporally evolving preferences in transformative contexts.

So it is not rational, when deliberating about parenthood, to agonize over which probability distribution over the known subjective outcomes you prefer in your set of probability assignments, because at this stage you don't even know the subjective values of the outcomes, and so you don't know which assignments of probabilities should be in your set and which should be excluded, and finally, you don't know what your new preferences will be if you undergo the change.[30]

[30] So the case is not analogous to the rational indecision described in Moss's "Credal dilemmas."

That said, it *would* be rational to agonize and deliberate if agonizing would give you information about which values you should assign to subjective outcomes. What sort of information would we be trying to discover by such agonizing? We'd be trying to discover what sorts of experience-based information we might already have available to us, given our previous experiences, to use in simulating the outcomes involving new experiences and assigning them subjective, phenomenal values.

The kind of evidence we'd need is not evidence that summarizes testimony or statistical facts about well-being. What we need is evidence that could tell us about which higher-order phenomenological features of the experience of what it's like to have a child are shared by experiences that prospective parents have already had or could have before making the decision to have a child. Such experiences are not experiences like changing a diaper.[31] They would have to be experiences that were able to teach you something about what it is like to be attached to your newborn, with all of its attendant emotion, vulnerability, stress, and excitement. This raises an interesting possibility. Perhaps the sort of experiences you'd need to have to grasp partial information about what it's like to have a child does not need to have anything (first-order) to do with children at all. The relevant experiences just need to share the right higher-order phenomenological character.

This is where empirical psychological work could contribute key insights. One important way empirical work could be of help is in collecting further testimony from parents and nonparents to try to determine the range and symmetry or asymmetry of possible subjective values. But a potentially more interesting and valuable approach

[31] They might include caring for the children of others, but only if something about experiencing the attachment you'd form to these children was relevantly like the experience of the one you'd form with your own child—and the only reliable way to know whether this is the case is through empirical information about such shared experiential character.

could be modeled on the idea developed in the next section, that we might be able to exploit hierarchical Bayesian modeling techniques to uncover shared higher-order phenomenological characters, for example, between concepts had by congenitally blind subjects and concepts had by sighted subjects. If psychological research could uncover shared higher-order phenomenological characters between experiences had by non-parents and experiences had by parents, this could be used to try and simulate features of the relevant outcomes to determine partial subjective values.

If you could get partial information about the value of what it would be like to have a child, you could use this to determine facts about the value space, to help you get a sense of how your preferences might change, and perhaps to determine which models, if any, for decision under ignorance could employed to make a rational choice.

What about credences? Here, again, empirical work can be of use. Once you have information about the subjective values of the outcomes, even partial information, you can use it to determine *something* about what your credences should be, even if what you determine suggests only a slight shift in probability assignments, or only excludes a few probability assignments from your set of options. As I noted above, even a slight shift in credences in a high-stakes case might be enough to help us rationally choose how to act, and if we could combine this information with models for value imprecision, I see the potential for getting some traction here. This is where agonizing about the decision might make more sense.

So what is key is discovering what sort of cognitive phenomenological information is relevant, that is, discovering, through empirical work, which parts of a person's previous experience, understood from her subjective point of view, could be exploited to give her the information or the ability she needs to determine the contours of what it is like to have a child. Such phenomenological information would be higher-order in the sense that it would be knowledge of relevantly similar or isomorphic features of her experiences that she could then use to cognitively model, or at least cognitively guide and constrain,

her assessment of the relevant subjective values. In the next section, I will discuss this in more detail.

Hierarchical Bayesian Models

Imagine a case where a congenitally blind adult must decide whether to have retinal surgery in order to be able to see. Such an adult, let us assume, has built his life around his blindness, choosing a career (he is a saxophone player, whose soulful music reflects his lived experience and his highly trained auditory capacities) and a way of living and understanding the world through touch and sound, a way of living that is deeply tied to his blindness.

Should he have the surgery? To decide, he would need to find a way around the epistemic wall created by the transformative nature of the choice, one that preserves a role for his first personal perspective in evaluating and assessing his subjective values and evolving preferences. Is there such a way? There might be.[32] Instead of simply eliminating the first personal perspective from decision-making and attempting to replace it with third personal testimony or descriptions, perhaps he could use empirical work to help him leverage knowledge from his previous experience to construct a partial phenomenological guide to the possibilities the surgery raises for his subjective future.

The idea for how to do this comes from work in cognitive science on learning and representation. This work explores how humans can make accurate and flexible predictions and inferences about novel situations using inductive inferences that draw on abstract or higher-order similarities between situations they've already experienced and novel ones they haven't. The models use probabilistic

[32] The connection that I sketch in this section between transformative decision-making and hierarchical Bayesian modeling was suggested to me by Josh Tenenbaum. I am indebted to him for very helpful discussion about the ideas developed in this section, in particular, for discussion about higher-order structure and the role of Bayesian models in learning and planning, and for further discussion about the adaptive value of discovery.

frameworks to provide explanations of learning and generalization in terms of Bayesian inferences. They exploit the fact that our cognitive phenomenology represents the way the world is in a range of different ways, and part of that representation involves an experience of the world as having a certain kind of structure. We use our experience of the structural form of the world to organize and guide our first personal understanding of future experiences, by using our previous experience of the relevant higher-order categories of the world to make predictions about our responses to new situations organized by those same categories.

Consider the choice to try a new species of grape. Assume that you've had lots of grapes of different kinds before, and in general, you are a big fan of grapes. They taste good. Should you try the new, bright orange grapes at the farmer's market? It seems like you should, reasoning that, in the past, you've liked red grapes, green grapes, blue grapes, and black grapes, that is, you've liked grapes of different colors, so you have liked grapes in general. While you like and appreciate the distinctive flavors of each different type, you also really like *grapes*, that is, you like this type of fruit along with liking the different instances of each member of this type. Liking a type of fruit, grapes, can be understood as liking something that all of the different types of grapes have in common. They share certain features, that is, they are similar in certain respects to each other. This sharing can be understood as sharing a universal, *grapehood*, or, equivalently for our purposes, as sharing a kind of abstract or higher-order categorical structure. Your experience leads to you think that you like things in the *grape* category.

When you reason that you'd like the new kind of grapes because you've liked grapes in the past, and so you decide to try the new grapes, you can make the decision by assessing the subjective value of eating grapes. To assess the subjective value, you evolve your first personal perspective forward to the outcome where you taste the new grapes, imaginatively constructing an outcome where the orange grapes have the same tasty properties as grapes you've had in the past, in virtue of

their membership in the grape category. Given your past experience, you assign that outcome a reasonably high expected positive value. (In real life, you probably do this implicitly rather than in any explicit way.) When you reason like this, it can be represented using computational modeling, where a higher-level categorization guides your reasoning about new lower-level cases (the new kind of grapes) based on how you'd reason about the higher-level features that the new case shares with your previous experience.[33]

The grapes example brings out how, if you can identify which higher-order properties are relevant to the cognitive assessment of a new outcome, you can draw on your previous experience of these higher-order categorical properties in other, ostensibly quite different situations in order to construct a cognitive model of the new situation.

In order to use this to partially model a transformative new experience, think of the type of cognitive modeling you'd need to do as employing knowledge you have about higher-order structure to generate new knowledge about the transformative experience. You take the higher-order structure that you abstract from sets of past experiences and use it to map out higher-order features of possible future experiences, much like you might lay a sheet of rice paper over a picture to trace its outline and then use it to generate a new drawing with the same form, but with different colors, different media and

[33] This approach can be framed in terms of hierarchical Bayesian models (HBMs). A distinctive feature of HBMs is that they can be understood as flexible, responsive tools for learning, and that only a few experiences are needed for the human cognitive system to be able to recognize and exploit higher-order categorical similarities. You only need to try a few grapes to grasp what grapes of that kind are like, and the more grapes of different types that you have, the more you can pinpoint and refine the higher order structure that they all share, and that you should use in your inferences about grapes of a new kind. Experience matters, and so you continuously update in response to new data. I'll come back to this point below when I discuss the long-term choice of raising a child. For relevant reading see Tenenbaum, Kemp, Griffiths, and Goodman (2011) "How to grow a mind: Statistics, structure and abstraction," *Science*, 331 (6022): 1279–85.

in a different setting. The relevant higher-order experiential facts are what should remain the same across the modeling of the very different, lower or first-order experiences.

This fact about human cognition can be exploited in certain cases of epistemic transformation. Consider the way we reformulated the choice to try a durian from chapter 2. If you've never tasted a durian, you don't know what it will taste like, and people have very different reactions to it. However, you might reason as follows: I like the way grapes taste, and given that fact, I like the way fruit tastes. Since I like fruit, I should try a durian. Here, you are drawing on your experience of grapes, which is experience of something with the property of *being fruit*, and evaluating the novel outcome of tasting a durian under the assumption that this outcome, because it is an experience of tasting something with the same higher-order property, *being fruit*, will have an experientially similar character.

There is something about this procedure which is clearly right, and which connects to the importance of using the subjective perspective to assess future outcomes. We should draw on our previous experience to make inferences about future outcomes, and while the outcome might be transformative at the level of our experience of the first-order properties, for example, at the level of like *what it's like to taste durian*, it might not be transformative at higher levels. If so, we should be able to draw on our experience of higher-order properties such as *what it's like to taste fruit* in order to partially assess the subjective value of the outcome of tasting durian.

Of course, since the shared feature between the experience of tasting grapes and the experience of tasting durian is the higher-order feature of the experience of tasting fruit, the contribution of this experience to the model for the overall subjective value of tasting durian will only be partial: you won't know much about the details of the experience, just something about its general character.

One problem which the durian example brings out is that just knowing about this sort of general, higher-level structure, even if there is no higher-level transformation, might not help enough. This

is because the higher-order structure or features of the outcomes may not contribute enough to the overall subjective value of the novel outcomes to matter. As we discussed in chapter 2, the response to the taste of durian varies widely, even amongst those who (usually) love fruit. In the durian case, it seems that the subjective intensity of the lower-order properties about the particular taste and smell properties of the durian sometimes swamps the contribution that its generically fruity features make to the experience.

This brings out how there might be many ways in which one's experience of higher-order features are not a useful guide to the subjective value of the novel outcome. Perhaps the experience of the higher-order structure is simply not the experience we notice when we assess the value. Perhaps we do notice it, but its value is swamped by the value of our experience of the lower-order properties. Perhaps other higher-order properties of the novel experience are the properties that determine the subjective value of the novel experience, and so on. Only if we draw on the *right* experiences of higher-order features when assessing the novel experience, and those experiences of higher-order features *matter* to the subjective value of the novel experience, can we exploit this feature of cognition to make progress on modeling outcomes of transformative choices in a way that is subjectively useful.

The problem derives from the fact that, while we can use higher-order cognitive modeling to assess the higher-order subjective value of novel experiences, because the experience is epistemically transformative, we cannot know which, if any, of the experiences of higher-order features that have contributed to the subjective value of previous experiences will carry over to the values we assign to outcomes of the epistemically transformative experience. (At least, we cannot know on the basis of our personal experience alone, because we don't know what the new experience will be like.)

A different issue arises when the experience is transformative at higher levels, or is transformative in a way that confounds prediction. The durian case is low-stakes, and so is not personally transformative. But individuals can experience preference change when they

undergo experiences that are simultaneously epistemically and personally transformative. If the decision-maker's preference changes are epistemically inaccessible to her, then there is a second reason why she may not be able to determine which higher-order features she should rely upon when assessing subjective values of future possible outcomes: which higher-order features matter to her assessment may depend on how she responds to the experience.

These problems bring out how, if an individual cannot know what it is like to have the transformative experience, before she has the experience, she cannot know, from her experience alone, which higher-order similarities between her future experience and her previous experiences are the ones that matter.

But what she *can* do in this situation, given that, in principle, we think we should be able to use knowledge from previous experience to model novel experiences in cases involving transformative experience, is use empirical work from psychology and cognitive science to try to determine the right higher-order features to use when making top-down inferences about these novel subjective outcomes.

In other words, perhaps we can use empirical findings about first personal values and preference changes to help us determine *which* higher-order features from *which* experiences, if any, we should use when, as individuals, we cognitively model our different possible outcomes for our acts involving transformative experiences. We can then use empirical findings and Bayesian modeling techniques, not to eliminate the subjective, first personal perspective, but to guide our understanding of how to selectively draw on our previous experiences to construct and evaluate possible outcomes for our subjective futures.

How might this work? There are many models for inductive learning that explore how higher-order structure is applied to discover facts about new situations. One promising approach, as I've suggested, uses hierarchical Bayesian modeling,[34] to find relevant

[34] Tenenbaum, Griffiths, and Kemp (2006) "Theory based Bayesian models of inductive learning and reasoning," *Trends in Cognitive Sciences*, 10 (7), 309–18.

abstract features drawn from our past experiences to probabilistically model our future ones. The Bayesian approach is especially useful for modeling how we can make strong inductive inferences from very sparse data and learn rapidly and flexibly. Such models for learning could be used in conjunction with data drawn from work on individuals who have had the transformative experiences to help identify the higher-order structure that is relevant to the forward mapping.[35] Employing cognitive models for inductive learning, where the right inductive constraint is empirically determined by finding the relevant structural similarity at the cognitive phenomenological level, could provide the first step towards constructing a new way to meet the normative rational standard for transformative decision-making.

The key is finding the right "overhypothesis": that is, finding the right higher-order structure to use to map forward to the phenomenologically new outcomes. This is related to the problem that has surfaced elsewhere when we wanted to use empirical information to guide us through our responses to transformative experiences, for example, when we wanted to use survey data to help us decide to have a child in chapter 3, and in this Afterword's sections on informed consent and the fundamental identity problem.[36]

The beauty of the work on hierarchical Bayesian modeling is that it shows how humans are actually able to *discover* or generate

[35] "The hierarchical Bayesian approach shows how knowledge can be simultaneously acquired at multiple levels of abstraction. . . Hierarchical Bayesian models (HBMs) include representations at multiple levels of abstraction, and show how knowledge can be acquired at levels quite remote from the data given by experience." pp. 307–8 of Kemp, Perfors, and Tenenbaum (2007) "Learning overhypotheses with hierarchical Bayesian models," *Developmental Science*, 10 (3): 307–21.

[36] "In Bayesian epistemology, *the problem of the priors* is this: How should we set our credences (or degrees of belief) in the absence of evidence? That is, how should we set our *prior* or *initial* credences, the credences with which we begin our credal life? David Lewis liked to call an agent at the beginning of her credal journey a *superbaby*" (Pettigrew (forthcoming) "Accuracy, risk, and the principle of indifference," *Philosophy and Phenomenological Research*). Knowledge of Lewis's "superbaby" phrase comes by way of testimony from Alan Hájek.

overhypotheses from very few experiences and then go on to use these overhypotheses to manipulate and understand their environments. In other words, research in cognitive science on probabilistic models for cognition and learning shows how incredibly good humans are at determining the right hypothesis to use, once they have just a little bit of evidence.[37]

This means that, when assessing future experiences, because of the levels of structure involved, we may be able to discover the relevant overhypothesis (one drawn from a higher-order level) given a combination of some data, layers of increasingly abstract structure, our ability to learn inductively from sparse data, and the existence of phenomenal similarities between the right categories of experiences. Assumptions about hypotheses at even higher levels are still necessary, but the hope is that the relevant nth higher-order hypothesis could be discovered from a combination of first-order experiences, the right n+1st higher-order hypothesis, and induction.

For example: you might assume *I like to try new foods* as a hypothesis, and, by combining this with some first-order experiences at different restaurants, along with your experience of liking grapes of different types, discover that, in particular, that you don't especially like trying new food of just any type, for example, you don't really like trying new fish-based dishes. What you really like about trying new foods is the experience of trying new *fruits*. You can then take *I like to try new kinds of fruit* as your overhypothesis, and use it to projectively model the outcome of trying a durian for the first time.

So the suggestion is that we could use higher-order cognitive phenomenological similarities to model and partly value new phenomenological outcomes, and in particular, use Bayesian learning techniques to discover which higher-order cognitive phenomenological similarities are the right ones to employ for this task. How

[37] Griffiths, Chater, Kemp, Perfors, and Tenenbaum (2010) "Probabilistic models of cognition: exploring representations and inductive biases," *Trends in Cognitive Sciences*, 14 (8): 357–64.

might such a strategy help in our high-stakes cases of transformative experience?

Consider our congenitally blind saxophonist. Should he have retinal surgery, if the choice were available? Should he choose to become sighted? Part of the choice must certainly involve consideration of the question of whether the subjective value of gaining a new sensory capacity and discovering the new experiences it provides is higher than the value of keeping things on the phenomenally same track.

To assess the subjective value of the outcome of becoming sighted, the blind saxophone player would need to use higher-order structure to move cross-modally from a one familiar sensory situation to a different, unfamiliar sensory situation. Even though the experience of being blind is very different from the experience of being sighted, some of the musician's previous experience might have the right sort of higher-order structure to allow him, from his own cognitive phenomenal perspective, to trace the experiential form and use it to model possible outcomes involving the experience of being sighted. If there is indeed any higher-order structure that would be shared by the musician's experiences before the retinal surgery and his possible experiences after the surgery, and if there is a way for him to identify and know that structure, he could use it as a partial guide for evaluating how his life could change.

The challenges here are significant. The first challenge involves the question of whether there *is* any relevant higher-order structure or form shared by the lived experience of blind adults and sighted adults that could give the saxophonist relevant knowledge about what it is like to see, such that he could evaluate possible outcomes of the decision to have a retinal transplant. Some of the values for these outcomes could depend, crucially, on the subjective value of what it would be like for him to see. Moreover, his ability to determine the expected value of choosing the surgery depends on his ability to determine how he'd value its possible outcomes given any relevant changes in preferences. The problem is difficult, for it is unclear what sort of means we have to determine the candidate higher-order structure that he could use.

What features of his previous experience could he draw on to make the needed assessments or to project his possible preferences?

Neuroscientific research suggests that the experience of being sighted is radically different from the experience of the blind, for blindness causes neural reorganization, and there are significant overall differences between the visual and occipital cortexes of blind and sighted individuals. And intuitively, it might seem that the congenitally blind, at least those with little or no visual capacities, must differ from the sighted with respect to their concepts of visual or action terms like "to show" or "to run," as well as of color predicates like "is purple," given the differences in the character of their dominant sensory experiences. Many theorists, dating at least from Berkeley, have agreed.[38] The view might seem to be confirmed by detectable differences in neural activity between blind and sighted individuals when they are performing similar linguistic tasks.[39]

But new research (Bedny et al.) suggests that the sensory visual experiences of normally sighted adults and the sensory experiences of the blind, including the congenitally blind, while deeply different

[38] Berkeley, *An essay towards a new theory of vision*. Or, for example, see Adam Smith: "Colour, the visible, bears no resemblance to solidity, the tangible object. A man born blind, or who has lost his Sight so early as to have no remembrance of visible objects, can form no idea or conception of colour. . . But though he might thus be able to name the different colours, which those different surfaces reflected, though he might thus have some imperfect notion of the remote causes of these Sensations, he could have no better idea of the Sensations themselves, than that other blind man, mentioned by Mr. Locke, had, who said that he imagined the Colour of Scarlet resembled the Sound of a Trumpet. A man born deaf may, in the same manner, be taught to speak articulately. He is taught how to shape and dispose of his organs, so as to pronounce each letter, syllable, and word. But still, though he may have some imperfect idea of the remote causes of the Sounds which he himself utters, of the remote causes of the Sensations which he himself excites in other people; he can have none of those Sounds or Sensations themselves" ("Of the External Senses," Glasgow Edition of the Works and Correspondence, iii: *Essays on Philosophical Subjects* (1795)).

[39] Bedny and Saxe (2012).

at the level of experience, are nevertheless similarly structured at a higher cognitive level. In particular, seemingly vision-based concepts like "to run" had by congenitally blind adults seem to share higher-order conceptual structure with the concepts had by sighted individuals. The researchers conclude that:

> While the sensory experience of blind and sighted people is drastically different, behavioral and neuroimaging data show that conceptual representations of these two groups are strikingly similar. These similarities hold for conceptual categories that, in our view, are among the best candidates to show effects of blindness. Conceptual representations used to understand concrete words, categorize objects and actions, and think about the perceptual states of other people are not images of sensory experiences. Humans have a rich repertoire of abstract representations that capture the higher-order structure of their environment in terms of events, objects, agents, and their mental states.[40]

This research suggests that there may indeed be some sort of higher-order structure available for a cross-modal application of blind experience to sighted experience. Obviously, questions abound, given the empirical underdetermination of the facts. What exactly is the form of the structure shared by the blind and the sighted? Is it higher structure that could be exploited, in much the same way as we exploited higher-order structure about trying new foods to discover the value of trying new fruit? Is there a relevant n*th*-order phenomenal similarity that is somehow experienced in a way that could give the blind the ability to assign partial subjective values to what it is like to see?

If the relevant structure exists and can be discovered, the blind saxophonist could, in principle, use this research to locate and draw on his relevantly similar previous experiences to identify the right abstract structure for the assessment of possible future outcomes and to model possible changes in preferences. Once the right inductive guidelines were established, he could use his previous experience of outcomes with this higher-order structure to construct partial

[40] Bedny and Saxe (2012) p. 74.

models of novel visual experiential outcomes, assigning them partial subjective values, that is, values based on what he knows outcomes with the right abstract structure would be like. To the extent that the blind saxophonist can evaluate the cost and the value of having new sensory modalities, perhaps he could predict any relevant preference changes and map forward to outcomes involving what it would be like for him to experience a new sense modality while experiencing an altered auditory capacity. This would give him an empirically supported method to use when trying to assign values and make big decisions about radically new experiences.

Obviously, even if the blind saxophonist can model something that will give him information about the subjective value of the new sense modality, the value of *what it is like to see*, in its full first-order manifestations, will still be inaccessible to him. As a result, even with the new strategy in place, we still come back to the fact that the radically new experience involved in transformative choice is epistemically transformative and potentially personally transformative. What it is like to see will have a complex effect on the saxophonist, since in addition to changing the way he organizes and lives his life and his relationships with his family and friends, it will change his auditory and tactile experiences, which are likely to change many of his central experiences, including his experience of playing the saxophone. Presumably, these knock-on effects will also have a major impact on his preferences and so on the expected value, for him, of becoming a sighted individual. In the case of the blind saxophonist, then, to have a successful partial model that will be of any significant use, he must find structure that supports an assessment of what it is like to see that is deep enough to also give him the new knowledge needed to determine what his new preferences will be as the result of becoming sighted.

So the structure used to model outcomes of transformative experiences would have to be informative enough to tell him enough about what it is like to experience seeing, so that he can model his new preferences (which would form as the result of becoming sighted) when he evaluates the values of the outcomes from the choice to have

retinal surgery. Or the structure must be complex enough to build this change of preferences in without his explicit knowledge of them, so that when he models his outcomes, the structure guides him to the values for the outcomes that reflect both his new knowledge of what it is like to see plus his new preferences as a sighted individual. This means that we need an extensive and far-reaching, highly developed abstract structure, one informed by the facts about the deep cognitive structure that might be constant across sensory modalities as well as by facts about deep cognitive structure that might be constant across major personal changes.

In addition, we face the computational limits of the experiencer: that is, while we have impressive capacities to model future outcomes, the abstract nature of the similarities we must rely on to model radically new outcomes may pose computational hurdles when we try to model these cognitively complex outcomes. In particular, some particularly important but causally downstream outcomes might be caused by the first-order nature of what it is like to see, and so might be especially hard to model. In other words, because what it is like to see will have a complex causal effect on the saxophonist, with many causal implications, it may be computationally intractable for him to model the relevant outcomes with only abstract structure to guide him. Moreover, we may not be able to effectively predict or evaluate the changes in preferences that would occur solely because of the first-order qualitative character of what it is like for him to see. And so, in high-stakes cases like this, where the intensity and kind of the transformative experience changes one's personal preferences, there are still no obvious solutions. This does not mean that using the higher-order structure doesn't help us with transformative choices; it means that the help may only be partial. But partial help is better than no help at all.

The case of the blind saxophonist brings out just how challenging it would be to uncover the right sort of higher-order structure to use in our inferences, but the fascinating work by Bedny et al. also makes the strategy seem at least potentially workable. If it did work, we'd have an example of how we can use models designed for partial ignorance

when making a transformative choice—even though we don't know what the first personal "on-the-ground" experience would be like—as long as we know enough about its abstract structure to start to predict and assign values.

The value of discovering the relevant higher-order structures is not that such structures can tell us everything about what it is like to have a transformative experience, but that it gives us a way to make an empirically supported, rational *start* on the decision problem. It points us in a direction we might take to develop a rough and approximate way of making a rational transformative choice involving a radically new discovery about the character of an experience.[41] It highlights a way that cognitive science, by discovering information about the relevant higher-order hypotheses, could be of use to those who are contemplating transformative choices.

My discussion in this section suggests a blueprint for further empirical work. But in the absence of such empirical work, for those of us who want to decide right now, for example, whether to have a child, how are we to proceed? How can we rationally approach transformative decisions before the empirical work pointing to how to determine and employ the right overhypothesis has been done? As I suggested in chapter 4, reformulating the decision seems like the best option. One way to frame the deliberation is predicated on our ignorance of the values of the outcomes.

If you are a Bayesian, perhaps you should embrace a hierarchical Bayesian approach to transformative choices based on revelation. The goal would be to generate a highly abstract overhypothesis about the value of revelation that you could employ when making personal decisions. *If* you have the right sort of evidence about the safety level

[41] There are connections here to using hierarchical Bayesian models to represent features of Kuhnian paradigm shifts in contexts of scientific discovery and revolution. See Henderson, Goodman, Tenenbaum, and Woodward (2010) "The structure and dynamics of scientific theories: A hierarchical Bayesian perspective" *Philosophy of Science*, 77 (2): 172–200.

of your environment (that's a big *if*), the place to start when making a transformative choice might involve the generation and assessment of very general overhypotheses, such as the hypothesis *I like transformative experiences*, or the hypothesis that *I dislike transformative experiences*. If, on the basis of previous experience, you select the hypothesis *I like transformative experiences*, you might be able to assess this by consulting your previous experience of particular transformative experiences (trying radically new activities, going to college, and so on) together with an assessment of the safety and the stability of your local environment to determine whether *I like transformative experiences* is the right overhypothesis to employ. In this way, you might be able to draw from your previous experience to partly guide your transformative choices on the basis of the desirability of revelation.

Unawareness

Another effective strategy for transformative decision-making might be to look to alternative models for rational decision-making under severe epistemic constraints. Work on decision-making under conditions of extreme ignorance, described as situations where agents make decisions where they are unaware of the outcomes,[42] which include situations where they do not know the values of possible outcomes of their decision, may be of help here.[43]

Imagine a game between a talented chess player and a mediocre one, where it is the mediocre chess player's turn to move. The mediocre player knows there are moves he can make that will have desirable

[42] I am particularly indebted to Joseph Halpern and John Quiggin for discussion.

[43] Halpern and Rêgo (2009) "Reasoning about knowledge of unawareness," *Games and Economic Behavior,* 67 (2): 503–25; Halpern and Rêgo (2013) "Reasoning about knowledge of unawareness revisited," unpublished MS; Heifetz, Meier, and Schipper (2006) "Interactive unawareness," *Journal of Economic Theory,* 130 (1): 78–94; Quiggin and Grant (2013) "Inductive reasoning about unawareness," *Economic Theory,* 54: 717–55.

and undesirable outcomes, but he is unable to visualize the game more than one or two moves ahead, and he has not memorized any appropriate strategies he could use to predict outcomes from making particular moves. In an important sense, then, he knows there are moves he could make that will result in particular outcomes of arrangements of pieces on the chess board, say, five or ten moves later, but he doesn't know what these outcomes are. He cannot, by himself, map out the game tree more than one or two steps ahead, and he has no strategy that he can use to determine the outcomes for him, so he is unable to know the outcomes of the different moves he is choosing between.

So, the mediocre chess player knows there will be outcomes a few steps down the tree that result from his moves, but he doesn't know what those outcomes are. Thus, he cannot work out the different possible paths from his next move to those outcomes, and he cannot proceed by calculating the expected value of his next move in the usual way. In this situation, how should he make his move? That is, how should he move in order to maximize his chance of winning, given the extreme computational constraints he faces? Is there nothing for him to do but make a random guess?

No. Even if he can't work out the consequences of his acts, he has other facts at his disposal, such as what he knows in general about the merits of various board positions, whether he wants to play a defensive strategy or take a more aggressive approach, and the worth of various chess pieces (he knows the difference, say, between the usefulness of having a Queen versus the usefulness of having a Rook).

What the literature on decision-making with unwareness discusses is how an agent in a situation like that of the mediocre chess player can maximize the rationality of his moves using other types of information he has at his disposal.[44] So, for example, in the chess game, the mediocre

[44] "While an agent cannot make decisions based on facts that he is unaware of, it is clear that awareness of unawareness can have a significant impact in decision making" (Halpern and Rego, "Reasoning about knowledge of unawareness revisited").

player can act rationally if he acts with the right understanding of how he is not aware of various outcomes. That is, he can act rationally if he acts with the knowledge that there are outcomes that he is "unaware" of in the relevant sense, making his choice of move in a way that is constrained by this knowledge, while also being guided by any justified reasoning principles and general evidence he has at his disposal.

We can apply this approach to the problem of transformative decision-making. Compare the choice the mediocre chess player must take to the choice you are to make between the ability to taste and having an entirely new sensory ability. When approaching the decision of whether to trade in your sense of taste, the main thing to recognize is that you cannot make this decision based on what it is like to have the new sensory ability in place of the old. You know that possible outcomes where your ability to taste is replaced by your new sensory capacity exist, but because you don't know what it would be like to experience these outcomes, you cannot simulate them from your first personal perspective. Since you cannot know these outcomes, you cannot value them, nor can you assign probabilities to states that lead to them, nor can you eliminate particular probability assignments from your set of possibilities. Thus, you cannot determine the expected values of your alternatives.

If you could know, through empirical work, which of the higher-order experiential features of the sensory abilities you already have are relevantly similar to the new sense, you might be able to exploit this information to help you make your choice. But we are also assuming that you are the first person to get the new microchip, and so there is no empirical evidence like this to draw on.[45]

[45] You could also use the models for unawareness if you did have some empirical information from cognitive science about common higher-order structure, since even with that information you'd only be able to partially model outcomes. The empirical information, in addition to guiding the construction of cognitive models of outcomes, could provide support for using certain general principles about sensory abilities.

Ideas from the work on unawareness may give us a way to start to model decisions for an agent in this sort of situation. Like the mediocre chess player, you know there exist outcomes you are unaware of, and so you act accordingly, constraining your decision by what you know, in general terms, about sensory abilities. For example, presumably, the new sense ability will involve some special capacity to detect new secondary qualities of the world. It might enhance or detract from other sensory capacities like smell or sight. While you cannot imagine or entertain scenarios that can tell you what it would be like for you to have this new sensory ability, you can use your general knowledge of what sense capacities are like to guide your decision.

In this sort of epistemically constrained context of discovery, facts about, broadly speaking, the safety and stability of your environment become especially relevant. If you have inductive or other evidence that you are in an unstable or unsafe environment, you should follow a precautionary principle when making a decision under unawareness, which suggests that you prefer acts that lead to known values and preferences.[46] If you have evidence that you are in a stable and safe environment, or evidence that mistakes can be easily corrected, you should follow an exploratory principle, which suggests that you prefer acts that can lead to the discovery of new, previously unknown values and preferences.[47]

Once we see what we'd need to do to reformulate and constrain the decision in order to make it rational, however, the epistemic poverty of our situation is laid bare. The proposal is to be guided only by the broad contours and constraints of the decision space when making a decision, that is, proceeding with almost no information or awareness of the qualitative nature of the particular unknown outcomes.

[46] Grant and Quiggin (2013) "Bounded awareness, heuristics and the precautionary principle," *Journal of Economic Behavior and Organization*, 93: 17–31.

[47] Given that learning new information can have significant value for causally downstream decisions, having a new experience can be of high value.

So, going back to your decision about sensory abilities, what would life be like without the taste of wine, ripe peaches, or French cheese, yet with a radically new sensory approach to the world? Who knows? You cannot make your decision on the basis of knowing what what life with the different abilities would be like. All you can use to make your choice are some very general, high-level facts about how you'd lose one kind of ability to experience the world and gain another, while the sensory abilities you kept might be altered in some way.

This takes us, then, back to the position developed at the end of chapter 4, where decisions are framed, not in terms of comparisons of the details of different ways of experiencing the world, but in terms of the value and the cost of revelation. When you make a transformative decision, what you assess is the value of revelation, that is, you choose between the alternatives of discovering what it is like to have the new preferences and experiences involved in the transformative change, or keeping the life that you know, and you can only make this assessment by relying on very general, and very abstract, facts.

Conclusion: Revelation and the Unknown

When we face a major life decision, we want to be able to consider what it would be like to experience the outcomes of each act we are considering so as to assess the expected subjective value of each act and compare it to its alternatives. But if an act we are considering involves a transformative experience, we must undergo the experience before we can know the subjective value of its outcomes. If the decision is momentous and irreversible, then we face a problem, for we cannot make the decision rationally by imaginatively considering possible outcomes and determining our preferences in the way we ordinarily do.

The problem cannot be avoided by relying on descriptions from those who have undergone the experience, because this will not communicate the relevant information about the nature of the experience. (If you've never seen color, you can't learn what it is like to see red just by having someone describe it to you.)

The problem cannot be avoided by relying on advice about what to choose from those who have undergone the experience, because you might not be similar enough in the relevant ways to respond to the experience in the same way as they did. You might not respond by having the same type of experience, or you might experience a different cognitive phenomenological intensity. Moreover, the experiences may actually change and form the preferences of those who undergo the experience.

The problem cannot always be avoided by relying on empirical data, for sufficient empirical data is often not available, and even when available may not be able to support the relevant counterfactual assessments. The problem cannot be avoided by eliminating preferences based on the first personal perspective, for to choose authentically and responsibly for outcomes involving your subjective future, you must consider your subjective preferences or you alienate yourself from your choice.

In sum, there are two kinds of problems with transformative decisions. First, we can't, despite the way the story is often told, approach transformative decisions by stepping back and evaluating our different subjective possibilities, imaginatively modeling outcomes and reflecting on the expected subjective value of our actions. In a situation of transformative choice, we simply don't know enough about what our lived experience will be like after the transformative experience. This has philosophical and practical implications for the way we live our lives. Second, we can't use normative rationality as a guide to navigate a transformative decision that is based on imaginatively assessing possible outcomes, because we lack the ability to assign values to outcomes or determine how our preferences might evolve.

Given this, we need to think differently about how we plan our futures, and about how, as rational decision-makers, we configure our decisions. As I've argued, we should draw on empirical findings when the right sorts of findings are available, and in this Afterword, I've discussed some promising theoretical and empirical ways to make inroads on the problem. But, crucially, in addition to managing

the decision-theoretic worries using more sophisticated modeling techniques, resolving the problem of transformative experience also involves valuing experience for its own sake, that is, for the revelation it brings.

Here, we connect back to the importance of subjective values, and how such valuing is distinct from merely valuing happiness or pleasure and pain. When we choose to have a transformative experience, we choose to discover its intrinsic nature, whether that discovery involves joy, fear, peacefulness, happiness, fulfillment, sadness, anxiety, suffering, or pleasure, or some complex mixture thereof. If we choose to have the transformative experience, we also choose to discover new preferences, that is, to experience the way our preferences will evolve, and often, in the process, to discover a new self. Or, if we reject revelation, we choose the status quo, affirming our current life and lived experience. A life lived rationally and authentically, then, as each big decision is encountered, involves deciding whether or how to undergo dramatic change to make a discovery about who you will become. If revelation comes from experience, independently of the (first-order) pleasure or pain of the experience, there can be value in having the experience of discovering how one's preferences and lived experience develop, simply for what such experience teaches. The game of life, then, is the game of Revelation, a game played for the sake of play itself.