1 Introduction

Do not worry about the future. If you come to it, you will come armed with the same reason that you apply now to the present.
– Marcus Aurelius, Meditations

New things happen to people all the time, and there is nothing surprising in that. But people also constantly make decisions about new things, and there is something strange about that. From tasting a novel dish to moving to an unknown city, from going on a date to going to war, how can people reasonably choose the unknown? A rational decision maker makes her choice by considering the costs and benefits of an outcome and weighing it against alternatives, and she can use simple probabilities to capture simple uncertainties. But what should she make of cases where she does not know the costs and benefits? And what should she do when the decision will alter her very core, the beliefs and desires she considers when making a decision?

Philosophical investigations of how people change their self in light of new experiences are not themselves new (Locke, 1700). Some of the oldest philosophy centers on this theme, such as Heraclitus' dictum: 'No person ever steps in the same river twice, for it's not the same river and they're not the same man' (Robinson, 1987). More recent philosophical work has explored both the difficulty of rationally reasoning about the self (Jackson, 1986; Parfit, 1984), and the fundamental difficulty of applying standard decision-making frameworks to big, self-altering decisions (L. A. Paul, 2014; Ullmann-Margalit, 2006).

In this chapter, we consider the psychological architecture that supports decisions about novel and transformative experiences, in light of advances in the computational modeling of thought. The first half shows how people can make informed decisions about trying unfamiliar things, by using rational inductive inference to evaluate novel experiences. This inference integrates previous experience, current preferences, and an understanding of how the world is organized, using a non-parametric hierarchical Bayesian model (Griffiths, Kemp, & Tenenbaum, 2008; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) to capture the structural uncertainty in such decisions. The model also accounts for higher-level decisions, by allowing for higher-order preference: people can decide to try something new because they like the novelty of it, regardless of the particular immediate sensation that experience brings.

The second half of the chapter tackles decisions about transformative experiences, meaning choices that can affect people's self, including people's decision-making faculty and the preferences it relies on. Reasoning about such choices requires that people have a theory-of-self, similar to the theory-of-mind they have of others. We present a formal framework for adjudicating between selves based on such a theory-of-self, and explore several alternative models within this general framework. We use both formal modeling techniques and empirical work to compare the different models within these framework, and end by considering what is still left out of a descriptive, computational account of making big decisions.

While we regard the following work as an important contribution to the philosophical and psychological issues around transformative experience, we consider the empirical results as initial explorations, and setting up the edifice for future work into attitudes towards real-world transformative experience.

2 Grape decisions: decision theory and novel experience

Imagine a philosopher and her friend the layman[1] walking through a street market. The layman comes across a yellow grape-like object he's never seen before, sitting on a counter next to red grapes, blue grapes, and green grapes. The layman picks up the yellow item, thinks for a moment, and pops it in his mouth.

Decisions like these are made every day with barely a second thought. Despite the ordinariness of it (or because of it), the philosopher decides to challenge her friend: How can the layman possibly have decided to try the new fruit, she asks, and claims this decision was made irrationally.

The philosopher argues as follows (the following is based largely on L. A. Paul 2014, 2015): In order for the decision to try a new experience to be rational, it must follow the rules of decision theory. In standard normative models for making decisions under uncertainty, a decision is made by considering all possible outcomes in terms of the likelihood they'll occur, and the possible benefits and harms to the decision-maker in case they do occur (see e.g. Weirich, 2004). The benefit and harm of an outcome can be captured by a utility function, and a rational decision is that which maximizes the agent's expected utility.

An agent's utility function may be informed by testimony (other people offering explanations, reasons, and self-reports about an experience), observation of others, or an agent's own personal experience. But, argues the philosopher, in certain situations neither testimony nor observation are enough to inform one's utility function (Lewis, 1990; L. A. Paul, 2014). And assuming neither observation nor testimony are adequate with respect to describing the experience of eating a yellow grape-looking thing, the agent apparently cannot assign a utility to the possible outcomes of the decision.

In short, making a rational decision where the utility depends on perceptual experience requires knowing what the resulting experience will be like. If the layman did not know in advance what the yellow item will taste like, he cannot rationally choose to try or decline tasting it.

The layman counters by pointing out that while he never tried this yellow grape-looking thing before, he has eaten green grapes, and purple grapes, and red grapes, and liked each of them. He reasonably assumed this new item was a sort of grape, and since he likes grapes, he thought he'd like this new one.

This yellow grape example (in a slightly different form) was previously discussed in the philosophical literature as an example of the difficulty novel experiences present for a rational decision maker (L. A. Paul, 2014). We consider it here in order to formalize the intuitions of the layman, and show that his decision is rational, as an example of the way previous experiences combined with a structured understanding of the world can be used to evaluate new items. At the same time, this formalization shows the machinery that underlies commonsense reasoning is not trivial. The layman may be justified in his decision, but the philosopher was justified in calling attention to it.

To spell out the layman's commonsense intuition more before formalizing it, it is an intuition based on an understanding of how the world is organized, and an understanding of the decision maker's own desires and preferences. That is, the layman abstracted away from particular instances (red grapes, green grapes, etc.) to a more general category (grapes). He then inferred that the new item must be a new sub-item of this general category, based on its observable properties (the yellow grapes look like grapes, so they probably are grapes). The layman further assumed that unseen properties of a new instance are similar to previous instances (a new grape will probably taste similar to previous grapes), and based on his previous preferences was able to infer his likely preference for the new item (since he liked the taste of previously tried grapes, he'll probably like the taste of this new grape). The layman's intuitive understanding is characterized in Figure 1.

[1] That is, a non-philosopher.

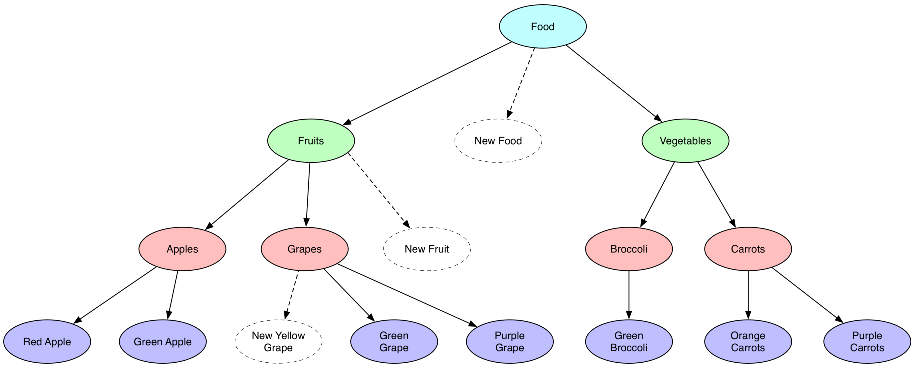

Figure 1: A simple intuitive model of the relationship between food items. Each instance of a specific food is at the lowest levels. While color varies within a given food, the shape and taste are less likely to do so. For instance, green and purple grapes share shape and taste, while green grapes and green broccoli share color but do not taste the same. Fruits on the whole have a similar taste, that is distinct from vegetables. A non-parametric hierarchical Bayesian model is able to hypothesize new categories and sub-categories at different levels of the tree.

The following section grounds the intuitive inference about simple novel experiences using a Bayesian hierarchical model. In order to account for the potential introduction of novel items, the model is non-parametric at every level .

2.1 Eating from the tree of knowledge: structured knowledge and decision theory

An agent's decision to try something new (such as a yellow grape-like thing) is partly an induction problem. The agent has experienced a number of objects before, and formed an associated preference over them that depends on their properties. The agent now needs to infer both whether a new object is a member of a previous category or a new category, and how the properties of the new object relate to previous objects. Our model is phrased in terms of food, taste, and shape in order to give a concrete example along the lines discussed in L. A. Paul (2014), but it applies more broadly to the decision to try or decline novel experiences with properties that depend on their category.

Our model posits a rich set of latent structure to perception that enables *unconscious inference*. Agents structure or process the "blooming and buzzing confusion" of their sensory data to make inferences about the hidden structure of the world (e.g. that there exists such a thing as a grape). Within such structures, agents are able to reason with much less information, and much more quickly, than if dealing with unstructured sensory inputs. As desiderata, a model that can accommodate a decision about a new item in a potentially new category should:

i Contain nested levels of domain specific information (similar to the structuring of grapes as fruits and fruits as foods).

ii Have lower levels inherit their properties from higher, more abstract levels. Thus, more similar categories should have similar prototypes, and also have similar variability in the ways they differ (apples and pears are more informative in making the decision about grapes, compared to broccoli and jalapenos).

iii Enable rich covariance relationships between features. For example, some features should be correlated with one another (such as the fruit's shape and its taste), while others should not (such as the fruit's color and its taste).

iv Accommodate for novel experiences by enabling a potentially infinite number of items at each level of the structure (this enables a simultaneous inference over whether a new item is a new grape, a new fruit, or even a new type of food).

Note that this list is incomplete, and a model that meets it is a simplification and not meant to capture the full range of decision-making and evaluation when assessing a new item such as a new food. In particular, the list does not mention compositionality of traits (and see Gershman, Malmaud, and Tenenbaum 2017 for a recent treatment of structured utilities that compose properties in the food domain). One model that does meet all the above criteria is a hierarchical Bayesian non-parametric model. We next define such a model in detail.

2.1.1 A Model for Choosing a New Item Based on Past Experience

Object representation The model represents a specific food object F_i as an tuple containing discrete taste, shape, and color attributes:

$$F_i = \{t_i, s_i, c_i\}, \tag{1}$$

where the different attributes are Boolean vectors. The taste attribute t_i is a Boolean 5-vector representation of each of the five major taste buds: sweet, sour, salty, bitter and umami. For example, a food F_i that is only sweet and sour will have $t_i = (1, 1, 0, 0, 0)$. Also, s_i is a vector representing the food's shape, and c_i is a vector representing color.

Utility function According to the model, agents have a utility function U that assigns a scalar value to a food object F. This utility function is informed by previous experience, and encodes the agent's expected *hedonic* pleasure or pain (although this may differ from the actual moment-utility, and see Bentham 1996; Kahneman et al. 2003; Kahneman, Wakker, and Sarin 1997; Loewenstein and Elster 1992). We assume for simplicity an agent's basic preference for a food item depends primarily on the taste of the food. An agent's utility function U is thus a mapping from the 5-vector t_f to a number, such that for two food items A and B, if $U(t_f(A)) > U(t_f(B)$ then the agent prefers A to B. The utility of objects not previously experienced can be reasoned about probabilistically, by inferring the likely t_{new} of the unknown object F_{new}.

In the running example, the layman derived utility from an item being similar to the grapes he's had before. Thus, we define a simple utility function:

$$U_{food}(F_i) = \sum_j U_{taste}(t_{ij}) \tag{2}$$

Where U_{taste} is the utility derived from the specific taste in the taste-vector. While we restrict ourselves in this section to utilities that depend on taste and sum similarly over the taste components, other utilities are possible, including non-linear combinations of taste and more abstract utilities such as a preference or dislike towards new and unknown items. Our goal here is not to accurately model the ways in which sub-utilities combine for taste. We consider these more general utilities in the discussion.

Generative Structure of Objects and Properties The specific food object $F = \{t_f, s_f, c_f\}$ is an instance at the end result of a generative category hierarchy, going from 'food' to 'category' (e.g. fruit) to 'sub-category' (e.g. grapes) to 'instance' (e.g. Concord grapes).

The top-most level, 'food', contains hyper-parameters used for drawing new 'category' objects. New categories are sampled in the following way, using the 'food' hyper-parameters :

$$\begin{aligned} t_{ij} &\sim Beta(\beta_{food}) \\ s_i &\sim Dirichlet(\alpha_{food}) \\ c_i &\sim Dirichlet(\alpha_{food}) \end{aligned} \tag{3}$$

Where t_{ij} refers to a specific taste component j within a taste vector related to food F_i. The 'sub-category' objects are sampled in a similar manner to the 'category' objects, though instead of β distributions over the set of possible tastes, a *Binomial* function represents variability in each taste vector within different instances of the same food 'sub-category'. This approximately captures the extent to which foods from each category are similar (how much we should expect grapes as a whole to be similar or different in taste). In our model, we specify all of these parameters within a reasonable range to describe the structure of the world before trying a new item, but they can be learned.[2]

The particular taste, shape and color attributes of a new food are sampled from its sub-category via from the multinomial and binomial distributions concentrated with the parameters of the 'sub-category' before it. These θ variables indicate the feature distributions across each food instance (e.g. yellow grapes and green broccoli).

We instantiated the above model in a *probabilistic program* through the WebPPL probabilistic programming language.[3] This generates a set of food instances with shape, taste, and color attributes. This model assumes that the agent is able to solve the inference about where to place the novel food object in the tree. In the technical appendix we provide a method for how, given access to observable properties (i.e. the food's shape and color), an agent is able to infer where to place the new object within the hierarchy, and from there infer its hidden properties (i.e. the food's taste).

Everyday Predictions The model above generated fruitful predictions about the novel experience of trying the yellow grape. It was conditioned upon the agent having eaten green grapes and green broccoli before, and that the object under consideration was yellow and grape-shaped. Additionally, we conditioned the sweetness, umaminess, bitterness, and sourness of the green grapes and the green broccoli as not equal (independent of what the taste of each was). [4]

Taste and utility of new item Given that an agent can place the new food object in the hierarchy (see the technical appendix), the agent is able to form a hypothesis about the taste of the new food, namely that it will probably taste like a green grape, and will deliver a similar utility. From this, the utility of eating the new grape can also be calculated (see Figure 2.1.1).

In short, even with very little information the above model is able to infer the relation of new objects to previously encoded objects, and then produce an informed hypothesis about the likely perceptual qualities of the novel experience.

2.2 Grape Decisions: Discussion

Our model shows how an agent can make a justified decision about a novel experience, using previous perceptual evidence to form a hypothesis about the utilities of experiences they have not had. While the decision is justified, the computations involved are not trivial, and rely on the ability to structure the world, and to account for potentially new objects through non-parametric reasoning. Such everyday, intuitive decisions made through 'common sense' or intuition, when prodded by philosophical questioning, turn out to have a rich structure to them. This is similar to the observation that facial recognition is fast and intuitive, but being able to perform it does not mean people have explicit access to the underlying computations that are carried out by the visual cortex when we recognize our friend coming towards us (Kahneman, 2011).

Our model considered simple utilities, tied to the particular properties of an object instance (the taste of a food item). But the model can also consider more abstract utilities, relying on the non-parametric hierarchical structure of the agent's knowledge. For example, the agent may derive negative or positive utility from 'opening up' novel categories at each level of the model:

[2] See Kemp, Perfors, and Tenenbaum (2007) for an account of a similar model that learns these parameters. This is captured via learning a posterior distribution $p(\alpha, \beta|n)$.

[3] For an in-depth treatment on the topic and the philosophy behind it, see http://www.probmods.org

[4] For the 'category' inference, we had (.1 1 .5) for the α parameters on the generative side corresponding to variability in shape, color and taste.

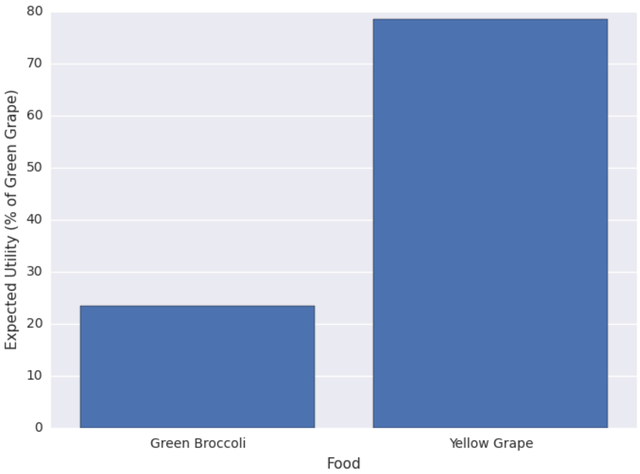

Figure 2: The expected utility of the new food object (yellow grape) and a known unfavorable object (green broccoli), relative to the utility of the favorable object (green grape).

$$U(F) = u_{taste}(F) + u_{new}(F), \tag{4}$$

where $u_{new}(F)$ is positive or negative depending on whether F is a new food instance. That is, regardless of particular taste experience, the very generation of a new experience category might in itself be associated with a positive or negative utility.

Consider the case of agent deciding whether to try a novel food called *durian* (L. A. Paul, 2014). Durian is a large and odorous melon native to Southeast Asia. By all accounts, eating a durian is a highly unique experience[5]. As testimony is unlikely to help describe the experience, except that the experience unique, how might previous experience inform an agent's reasoning about whether to try such a food?

The agent may derive pleasure (or pain) from creating a whole new category of fruit, or even a whole new category of food. Such pleasure from exploration (or aversion to it) may be related to the 'Openness' factor in human personality, as measured in the popular and empirically vetted OCEAN framework (McCrae & John, 1992). Our model, coupled with the non-parametric extensions in the appendix, extend this measure by representing a hierarchical notion of novelty itself. One can also imagine valuing novel experiences at higher levels of the hierarchy more, as they provide a more foundational change to one's understanding of the world. So, while a new object may have certain unknown ground features, and present a problem for a decision-making account that relies on knowing these features, the very fact that the object is new presents a higher-order, observable feature which the agent may take into account in its decision making. We mean this as a normative point about higher-order decisions can be made, and how they can be captured specifically as the opening of a node new in a hierarchical representation, and do not mean to suggest that descriptively all people will in fact weight novelty for good or ill in their decision making, nor that novelty is the only higher-order observable feature of new objects.

For all its uniqueness and novelty, the experience of eating a durian for the first time likely leaves the decision-maker essentially unchanged. This is an epistemic change, in that it opens up new possibilities of pondering experiences, but it is not a self-change to the agent's model of itself. A less common, far more gnarly, and much more interesting case is that of transformative experiences that change the agent itself in a more fundamental way. These are experiences where the desires, intentions, qualities, and beliefs that led to the decision are themselves altered. This is the case we turn to next.

[5] Paul writes: "One important chef says, "The only way to describe its taste is 'indescribable.'".. [other] descriptions people have [include]: "Eating vanilla ice cream by a sewer" or "French kissing a dead rat."

3 Who Decides on the Decider? Decision Theory and the Intuitive Theory of Self

How do we understand other minds? A leading view in cognitive psychology is that people construct an intuitive theory-of-mind when reasoning about other people (Dennett, 1989; Gopnik, 1993; Happé, 2003). According to this intuitive theory, other agents have beliefs, desires, intentions, and various other mental states that lead them to take certain actions. While these desires and beliefs cannot be directly observed, we can infer them from how people act, and use them to explain past actions, and to predict future actions. This inference process has been formalized in recent years as 'Bayesian theory-of-mind' (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Baker & Tenenbaum, 2014; Hamlin, 2013; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern & Kemp, 2015), and used to capture adult and children's reasoning about the social and non-social goals, beliefs, and relations of others.

How do we understand our own mind? A constructed notion of an essential 'true self' as the binder of intentions, desires, and in particular moral traits has been explored in cognitive psychology (e.g. Newman, Bloom, & Knobe, 2014; Strohminger & Nichols, 2014; Strohminger N. & Knobe, in press). But what would a formalization of an intuitive 'theory-of-self' look like? It may be quite similar to how theory-of-mind is constructed for other people (Baker & Tenenbaum, 2014; Saxe, 2009), informed by additional (though not full) access to memories and internal states. Note that on such a view it is not necessary that the theory-of-self truly aligns with the self, any more than the theory of intuitive physics truly aligns with real physics, or theory of mind aligns with other people's *actual* decision making, which it the subject of decision theory and behavioral economics. The deliberations and explanations that a person gives to themselves when contemplating a self-changing decision may be quite different from the actual reasons that drive them to make a decision.

How do we understand a transformative experience? Ullmann-Margalit (2006) defined transformative decisions as those that abruptly alter the core desires and beliefs of the decision maker [6]. A transformative decision is one that takes us from our current model of our self (as an agent with particular desires, intentions and beliefs), to a different agent (with new desires, intentions or beliefs, see Fig Xb). Transformations that happen with discontinuities of space or time frequently call into question a notion of self (such as the teleporation experiment, see Parfit 1984). To this we can add discontinuities in a more abstract space of beliefs and desires.

Formalizing the intuitive theory-of-self along the same lines as Bayesian theory-of-mind casts the deliberation about transformative decisions as an agent-based decision-making problem, and gives the ability to specify quantitative differences between different selves. That is, the deliberation can be seen as a problem of expected utility maximization, where the overall utility is defined over agents (current and transformed) that themselves have different utilities and beliefs. But such a formalization leaves open many questions about how the decision should be made: What overall utility function should be used to arbitrate between different agent-specific utility functions? Should the decision-maker construct an external meta-agent to arbitrate between their current self and their potential future selves? Should an agent's current utilities and beliefs matter more than the potential future self, and in what way? Can one even make a rational decision about a future self, in the same way that one makes a rational decision to eat a yellow grape?

These questions correspond to questions raised by philosophers examining transformative experiences (e.g. L. A. Paul, 2014)). Pettigrew in particular Pettigrew 2015 recently showed that a transformative decision can be 'rational', in the sense that it can be made within an appropriate decision making framework. In Pettigrew's framework, local utilities are assigned to successive chunks of time, and weights are assigned to each utility. The question then becomes that of which weighting one wishes to place on different utilities

[6] As Ullmann-Margalit puts it: "[L]et us think of cases of opting as cases in which the choice one makes is likely to change one's beliefs and desires (or 'utilities'); that is, to change one's cognitive and evaluative systems. Inasmuch as our beliefs and desires shape the core of what we are as rational decision makers, we may say that one emerges from an opting situation a different person. To be sure, there is a sense in which every choice changes us somewhat. The accumulation of these incremental changes makes us change, sometimes even transform, as life goes on and as we grow older. But what I am here calling attention to are the instances in which there is a point of sharp discontinuity."

(perhaps you care more about the utilities of your current self, or perhaps you discount utilities after a transformation). Pettigrew himself points out that this leaves open the question of how to set the weights, but the problem is at least rationalized. However, as Paul points out, this framework only treats the decision making problem from the point of view of the current self S_c (L. Paul, 2015). From the point of the transformed self S_t, the utilities will be assigned a different weight, and there is perhaps no reconciling the two. Imagine for example an artist considering becoming a lawyer. The artist imagines going to lawyer parties and assigns the derived utility a low weight, for various reasons. However, if she were truly to become a lawyer, from the perspective of the transformed self, these same parties might be assigned a high weight. One cannot then say that the artist is choosing rationally to not become a lawyer, from the perspective of the lawyer. But how is the artist to choose? Pettigrew considers placing second-order utilities over utilities, but rejects this as leading to an infinite regress of higher-order utilities. However, this worry should be embraced as part of the problem of transformative decision making. Furthermore, it may be that in practice people only consider a small or single transcendence in utility-functions.

In the following sections, we describe a common framework for agent-based decisions, and then formalize a simplified framework for a theory-of-self, along the lines of a model for theory-of-mind for other agents. We show the different possible models an agent may have for arbitrating between different selves. We evaluate these different models by considering their predicted output, and comparing it to the empirical behavior of participants in 4 experiments.

3.1 Agent-Based Decision Theory and Simple Decisions

We assume that people explain the behavior of others by thinking of them as agents acting rationally to achieve their goals, subject to their beliefs about the likely state of the world Baker, Saxe, and Tenenbaum n.d.; Jara-Ettinger, Gweon, Tenenbaum, and Schulz 2015. Again, this view is agnostic as to the mechanism people *actually* use to make decisions. Rather, this is a formalization of the intuitive theory people use to explain and understand, to others and themselves, the causes of their decisions.

Following standard planning algorithms (see for example Russell & Norvig, 1995), we assume a world that is in a particular circumstance c out of a set of possible circumstances C[7]. At each point in time an agent can take an action a out of a set of actions A, the result of which is a new world circumstance $c^{'}$. A transition function T defines the probability of moving from one circumstance to another circumstance, given an action:

$$T(c, c', a) := P(C_{t+1} = c' | C_t = c, A_t = a). \tag{5}$$

Agents derive value from achieving their goals in a given world circumstance $C = c$, or (possibly ordered) set of world states $c_i \in C$, which we can represent through a utility function $U(c_i)$.

An agent does not necessarily have direct access to the true circumstance of the world, or to the true transition function. Rather, the agent can make observations of the world $o \in O$, and use the observations to inform and adjust their belief about the current circumstance. An agent's belief about the current circumstance, and the likely transition function given those observations, are represented via functions $B(C|O)$ and $B(T|O)$.

Agents have a planning procedure, $PLAN$, which takes in the agent's utilities, beliefs, and possible actions, and returns a probability distribution over the set of actions the agent can take:

$$PLAN(U, B, A, C, O) \rightarrow P(A = a | U, B, O = o), \tag{6}$$

where $P(A = a|U, B)$ is the probability of taking a particular action a, given utilities U, beliefs B, and observations o. A rational planning procedure returns a distribution over actions such that the agent acting in accordance with this distribution will maximize its expected utility. There are many different ways to implement PLAN, such as Markov Decision Processes, Partially Observable Markov Decision Processes (Kaelbling, Littman, & Cassandra, 1998), and Planning-as-Inference (Todorov, 2004).

[7] The common terminology uses 'state' instead of 'circumstance', but the variable S will soon be used to indicate different selves.

By assuming that agents implement a planning procedure that produces actions in order to maximize utility under constraints, an outside observer can use inverse Bayesian reasoning to infer the latent variables (the likely beliefs and utilities of an agent) given the observed variables (the agent's actions and the world):

$$P(U,B|A=a,C=c) \propto P(A=a|U,B,C=c)P(U,B), \tag{7}$$

where $P(U,B)$ is the prior belief the observer places on the beliefs and utilities of an agent, and $P(A=a|U,B,C=c)$ is given by a planning procedure as explained above. This is the basic notion of Bayesian Theory of Mind Baker et al. (2017). In what follows, we focus mainly on the representation of selves as decision-making agents, and less on the inverse-planning aspect of Bayesian Theory of Mind.

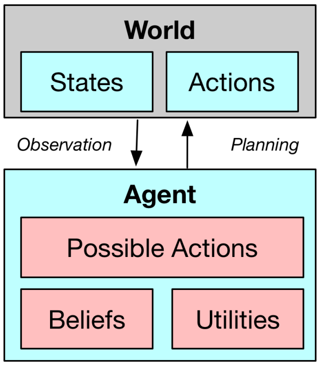

Figure 3: This is a simple model for an agent's intuitive planning procedure. Agent's beliefs, utilities, and possible actions receive inputs from observations they make about the world. This inform the actions an agent takes in the world.

3.2 Transformative Choices and an Intuitive Theory-of-Self

We can use the agent-based decision making framework described above as a starting point for an intuitive theory-of-self. The current self (S_c) is a decision-making agent with structured beliefs, goals, intentions, and other mental qualities that drive its actions. The potential transformed self (S_t) is also an agent, with possibly different beliefs, goals, intentions and mental qualities.

The problem of transformative experience can be recast as a decision making problem. Given the choice of staying as the current agent S_c or changing to a new agent S_t, what should a rational person choose? Note that in this formalization the different agents have different utilities, but this can be recast as agents placing different weights on the same utilities in Pettigrew's framework Pettigrew (2015).

On a naive "View from Nowhen" account[8], the agent can simply apply standard expected utility theory to this dilemma. That is, the agent considers the likely future circumstances it will encounter as S_c and S_t (different selves may encounter future circumstances with different probabilities), and compares the expected utility from these circumstances under the utilities of the current self S_c, and under the utilities of the transformed S_t. The agent then chooses the self who is most likely to be maximize their (expected) utility. In a sense, the agent would be constructing a meta-agent deciding on the meta-action of what agent to be. The meta-agent would map from the set of possible agents A to a particular agent. If we restrict ourselves for simplicity to two agents (current and transformed), we have:

$$MetaAgent(S_c,S_t) = \max_{S_i} E(U_{S_i}|S=S_i) = \max_{S_i} \sum_c P(c|U_{S_i},B_{S_i}) \cdot U_{S_i}(c), \tag{8}$$

where $U_{S_i}(c)$ is the utility agent S_i gets from the state of the world c, and $P(c|U_{S_i},B_{S_i})$ is the probability of the world being in state c, given that an agent is S_i. That is, $P(c|U_{S_i},B_{S_i})$ takes into account the particular beliefs and goals of S_i, and thus its likely actions and resulting state of the world. One could include other mental properties and qualities beyond beliefs and utilities, but these are the common ones explored in theory of mind research.

There are at least two major difficulties with this formulation. The first hurdle is the one explored in the first section - the difficulty or impossibility of imagining certain future states. As we saw, it is partly possible to get around this difficulty by considering that circumstances and experiences in the world are structured in nature, relying on past experiences and considering meta-utilities on things like novelty. Such a formulation does not fully solve the epistemic concern, but this difficulty is at any rate not our focus here.

[8] C.f. 'The view from nowhere' (Nagel, 1989). We thank John Schwenkler for suggesting this term

The second hurdle, which is our focus, is that even if it is possible to imagine future circumstances and utilities under a transformed self, it is unclear *whose* utilities should count. Under a simplistic view-from-nowhen account, such as that expressed in equation 8, it is be perfectly reasonable to accept the following suicide-for-happiness bargain: your current person is pulverized to a pulp, and the matter is used to reconstruct a new person, who is guaranteed to be happier than you in every way in every circumstance. Most people would reject this idea not because of a failure to simulate the new person, but because the new person is not them. In this suicide-for-happiness bargain, the destruction of the self is physical. But it may be that deep changes to beliefs, utilities, and other aspects of the self also count as a potential self-destruction. The utility of the stranger, the transformed self, does not matter to the current self, and what's more, the stranger is not the one currently making the decision. But an asymmetry is not apparent in equation 8.

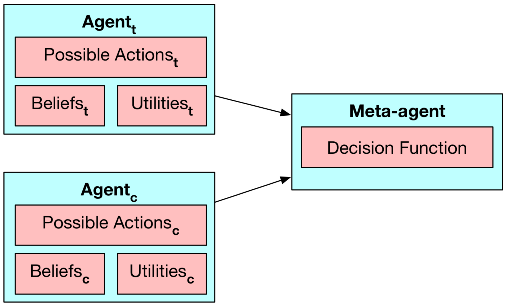

Figure 4: We can think of transformational decision are decisions a "meta-agent" makes reasoning about our current self and future selves.

What other view could people adopt, other than a view-from-nowhen? At least for preferences, one could construct a non-symmetric meta-agent for whom the preferences of the current model-of-self are taken more into account. Informally, in this model, as the agent calculates the expected utility of its new self, it asks 'what would my *current* self think of this circumstance?', even though their current self would no longer experience this circumstance. For concreteness, imagine an aspiring artist contemplating becoming a lawyer, as her demanding father wants. The artist knows that once she becomes a lawyer, she will take part in a standard lawyer's life and will probably even enjoy it. The artist pictures herself in a year's time going to a dinner party with other lawyers, and finds herself bored by the very idea. Alternatively, she imagines going to an exhibit opening at a small gallery, and is quite happy with the thought. The a-symmetry is that she is thinking of the dinner party from the point of view of her current preferences – the artist trapped in the lawyer's body. For the exhibition, she is not considering a trapped lawyer-self and their view on the circumstance.

The decision we consider is again made by a meta-agent, but a meta-agent that calculates EU_i with an eye towards the current self. More formally:

$$BiasedMetaAgent(S_c, S_t) = \max_{S_i} \sum_{c} P(c|U_{S_i}, B_{S_i})[(1-\alpha) \cdot U_{S_i}(c) + \alpha \cdot U_{S_c}(c)] \tag{9}$$

where U_{S_c} are the utilities of the current self. The parameter α controls the degree to which the utilities of the current self are taken into account compared to the utilities of other selves, and may be different for different people. If α is set to 0, we recover the view-from-nowhen in equation 8. If α is set to 1, the agent parochially considers future events only from the point of view of its current preferences, without taking into account the possibly different preferences of the new self.

Which of these views is more accurate in describing how people act? We first consider the difference between the view-from-nowhen (MetaAgent) and the view-towards-self-preferences (BiasedMetaAgent) through two studies that ask people to consider a change of their desire.

3.3 Study 1 - A Simple Increase of Utility

As a warm-up study, we asked participants to consider a choice between a pleasant and unpleasant experience. Before being asked to choose which experience they preferred to go through, they were given the option of increasing the degree to which they enjoyed the pleasant experience. We expected most people to make use of this option, as predicted by both a view-from-nowhen model, and a view-towards-self-preferences model.

3.3.1 Methods

Eighty participants were recruited on Amazon’s Mechanical Turk Service (56 male, 23 female, median age = 30, ranging from 21 to 70 years old, 1 participant failed a catch question and is not included in the analysis).

Participants were asked to imagine that they are facing two doors. If they open the first door, they will be given a mildly painful electric shock that lasts for five minutes and leaves no lasting effects. If they open the second door, they will be given a fluffy puppy to pet for five minutes. Participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 (*extremely unpleasant*) to 100 (*extremely pleasant*). After rating the two options, participants were asked to choose which door they preferred to open. Participants were then told that prior to opening the doors, they could press a button that would make the pleasant experience more pleasant (+20 points on the pleasantness scale, capped at a maximum of 100). All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press such a button prior to opening the doors. They then explained their reasoning, and provided their age and gender.

3.3.2 Results

Participants unsurprisingly rated the anticipated experience of petting the puppy as pleasant (+70), and the experience of the electric shock as unpleasant (-65) in all experiments. Participants indicated they preferred the puppy to the shock in all experiments. All participants went with their indicated preference. Most participants who went with their preference chose to press the button before doing so (71%, 95% CI = 62% - 80%). This result indicates participants are able to reason about their own simple experiential utilities, and are willing to change their preferences. We did not expect nor find any significant gender effects in this or the other experiments. This result is expected both under an unbiased meta-agent, and from a biased meta-agent. In this case, the increase in utility for the transformed self was in line with the preferences of the old self. It is interesting, however, that not *all* or even a large majority of participants were willing to change their utilities, indicating that participants may place value on having the specific utilities they currently have. Some comments were elucidating in this regard: “I don’t want it to be enhanced, it’s like taking a drug”, “I don’t need [the experience] to be any more pleasant”, “It’s already very pleasant”, and so on. We return to this point in Study 4. We next consider a reversal of utilities.

3.4 Study 2 - A Simple Reverse of Utility

3.4.1 Methods

We recruited eighty participants on Amazon’s Mechanical Turk Service (42 male, 22 female, median age = 30, ranging from 21 to 60 years old, 16 participants failed a catch question). Participants from Study 1 were excluded from participating in Study 2.

Participants were asked to imagine a situation similar to Study 1 (a choice between petting a fluffy puppy and receiving a mildly painful electric shock). As in Study 1, participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 (*extremely unpleasant*) to 100 (*extremely pleasant*), and to indicate their choice of door. Participants were then told that prior to opening the doors, they could press a button that would make the unpleasant experience even more pleasant than the pleasant experience (for example, if ’painful shock’ was rated as ’-20’ and petting the puppy was rated as ’+30’, the shock would now be as pleasant as ’+50’). All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press a button prior to opening the doors. They then explained their reasoning and provided their age and gender.

3.4.2 Results

In this experiment, 31 of the 64 participants chose to press the button prior to opening the doors (48%, 95% CI = 39% - 58%). The z-statistic of the difference between experiments 1 and 2 was 2.73, indicating statistical significance.

3.4.3 Discussion

As opposed to Study 1, only about half of participants are willing to press the utility-altering button. Such a result is not in line with an unbiased meta-agent, nor is it in line with a view maintaining that experiential utilities are best kept unchanged. The biased meta-agent (Equation 9) is in line with these results, with $\alpha \approx 0.1$. That is, the utilities of the current self matter, but to a small degree. Note that this analysis does not disentangle the possibility that the difference in decision-rules is happening at a population level. That is, it is possible about 90% of the population uses an unbiased meta-agent for such meta-preference decisions ($\alpha = 0$), while the other 10% uses an extremely current-biased meta-agent ($\alpha = 1$). It is also possible that some people are expressing a meta-preference for not having their preferences changed, regardless of the content of the preferences and regardless of simulating a future self from a particular view-point. This added component is discussed in Study 4.

3.5 Study 3 - A Change of Belief

So far we have considered changes of preference, and found evidence in support of people continuing to take their own preference into account even after a transformation of preference occurs. It is as if people are saying 'If I push the button, I will want to choose the shock. And I do not like being shocked'. But what of changes in belief? Do the beliefs of the current agent also mix in with the beliefs of the new agent?

In equations 8 and 9, it was necessary to compute $P(c|U_{S_i}, B_{S_i})$, the probability the world will be in state c given that the agent is S_i. This can be used to predict that if our future self believes a particular outcome is behind the left door, and if our future self likes that outcome, then our future self will open the left door. But it is possible that our future self is mistaken about the actual outcome of their action. Perhaps right now, our current self believe the left door is empty, or has something horrible behind it. Even if we imagine a future self with a different state of belief, we consider it a *false* belief (similar to the way we can consider false belief in others (Perner, Leekam, & Wimmer, 1987)). Beliefs might hold a different status than utilities for imagining future selves, in that we can imagine ourselves with certain arbitrary utilities (e.g. you like vanilla, but you can imagine yourself liking chocolate, without that preference being a 'false' preference), and we can imagine ourselves with certain beliefs, but we cannot hold those beliefs to be the actual state of the world (Shoemaker, 1995).

We next consider how people might take into account a different belief of a future self.

3.5.1 Methods

We recruited eighty participants on Amazon's Mechanical Turk Service (42 male, 22 female, median age = 31, ranging from 20 to 68 years old, 16 participants failed a catch question and are not included in the analysis). Participants from Studies 1 and 2 were excluded from participants in Study 3.

Participants were asked to imagine that they are facing two doors. If they open one door, they will be given a mildly painful electric shock that lasts for five minutes and leaves no lasting effects. If they open the other door, they will be given a fluffy puppy to pet for five minutes. Unlike Study 1 and Study 2, participants were told they do not know which door leads to which outcome. Participants were asked to rate the relative pleasantness or unpleasantness of the two experiences, ranging from -100 (*extremely unpleasant*) to 100 (*extremely pleasant*).

After rating the two options, participants were asked to choose which option they preferred to occur. Participants were then informed that prior to opening the doors, they could press a button that would make them absolutely certain the pleasant experience was behind the door on the left. It was emphasized that pressing the button would not reveal the location of the outcomes, but merely change the certainty of the participants. All effects of pressing the button were said to disappear after five minutes. Participants indicated whether they would press such a button prior to opening the doors. They then explained their reasoning, and provided their age and gender.

3.5.2 Results

In this experiment, 24 of the 64 participants chose to press the button prior to opening the doors (32%, 95% CI = 24% - 41%). The z-statistic between experiments 2 and 3 was 1.98, bordering on significance.

3.5.3 Discussion

The majority of participants rejecting the pressing a belief-altering button cannot be explained by an unbiased meta-agent that holds the beliefs and utilities of all agents equal. On the view-from-nowhen account, the expected utility for an agent that knows the location of a preferable option is strictly greater than one with uncertainty about the location of a preferable option. A simple preference-biased meta-agent that over-weighs the preferences of the current agent would also lead to pressing the button, and so also does not account for the result. It is also possible that people prefer not to have their beliefs and preferences manipulated at all regardless of the content, a point which we consider in the next section.

This result is possibly driven in part due to the fact that while people can conceptualize having different beliefs to themselves, as well as acting on those (presumably false) beliefs, they do not think that their actual beliefs regarding the transition function of the world is wrong (c.f. Shoemaker, 1995). Thus, if they currently believe something bad is behind door A, then they can imagine believing otherwise (that it is good), and they can imagine acting on that false belief (opening the door), but when predicting what would actually happen as a result opening the door, they maintain their current belief (something bad would happen).

This reasoning can be captured by adjusting the previous self-centered meta-agent to take into account the beliefs of the new self S_t for calculating the likely actions of the new agent, but to use the transition function of the current self S_c for calculating the actual outcome.

More formally, we can break up the term $P(c|U_{S_i}, B_{S_i})$ into:

$$P(c|U_{S_i}, B_{S_i}) = \sum_{action} P_{S_i}(c|c', action) \cdot P(action|U_{S_i}, B_{S_i}), \tag{10}$$

where $P_{S_i}(c|c', action)$ is the transition function of the world as agent S_i believes it to be, and $P(action|U_{S_i}, B_{S_i})$ is given by agent S_i's planning algorithm. Instead of using this term, it is possible to use:

$$P(c|U_{S_i}, B_{S_i}) = \sum_{action} P_{S_c}(c|c', action) \cdot P(action|U_{S_i}, B_{S_i}), \tag{11}$$

Notice the move from the transition function P_{S_i} to P_{S_c}. That is, S_i's actions are predicated on their (possibly false) beliefs about the world, but the actual result of their actions is dictated by what the current self believes is going to happen.

This is an agent that can imagine a future self being delusional and making bad choices (under the believes of the current self), but still imagines the result of those choices as happening according to the current self's belief. Such a self-centered meta-agent would be agnostic about pressing the button, as the expected utility for the agent is the same whether they push the button it or not.

Empirically, however, people were not agnostic between pushing the button and not pushing the button. There are several possible reasons for why this might be the case, including preference for inaction over action (Thaler & Sunstein, 2008). One possibility is that people reject changing their beliefs when there is no reason to do so acting on the 'principle of rational belief' according to which there should be a causal link between observations and beliefs (Baker et al., n.d.). Again, this would point to the asymmetry between changing utilities (with preferences being in some sense arbitrary) and changing beliefs (people have the beliefs they do for a reason).

3.6 Study 4 - A Jump in Self-Space

It appears people value having the beliefs and utilities they currently do, regardless of what the particular values of those are set to. There may be some inherit cost (or benefit) for changing from one self to a different self, not captured by the meta-agents considered so far.

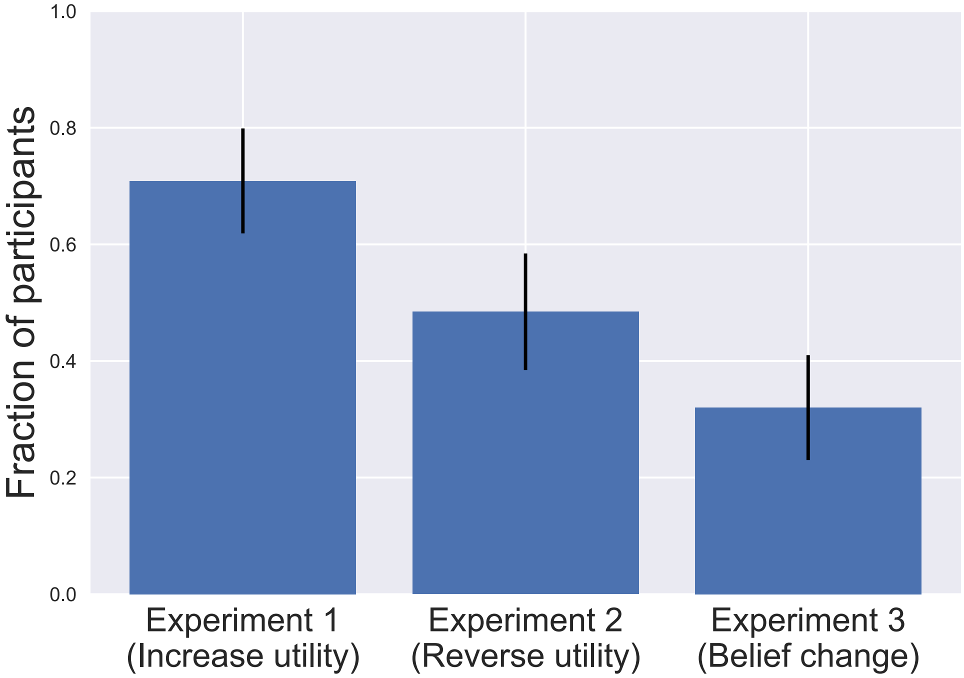

Figure 5: Fraction of participants choosing the transformative option in Experiment 1-3. Black lines indicate 95% confidence intervals.

To formalize this intuition, we can define a distance metric between selves, as the distance between the belief functions and utility functions of the current agent and the new agent. This distance can then be associated with a utility function (positive or negative). When faced with a decision about whether to change its beliefs and utilities, an agent also thinks about how much it is going to change *itself* in making the decision.

In general then, we expect:

$$DistanceMetaAgent(S_c, S_t) = \sum_c P(c|U_{S_i}, B_{S_i})[(1-\alpha)\cdot U_{S_i}(c) + \alpha \cdot U_{S_c}(c)] + U_c^d(d(S_i, S_c)), \tag{12}$$

Where $d(S_i, S_c)$ is the distance between the potential self S_i and the current self S_c. U_c^d is the utility that the current self S_c places on the distance d. There are many ways for defining a distance between functions, or between distributions. The exact implementation does not matter for the general argument of placing a utility over having a particular utility, a particular set of beliefs, or a trait.

In the next experiment, we consider changing the trait and utility of a person, and the resulting decision problem this poses. We assumed that decisions of changes to the self are partly driven by difference to the self and partly by utility, and so we examine positive changes to a quality while decreasing hedonic utility, and vice-versa. We predict people will not choose on the basis of increase in happiness alone, and that their choice will be related to the amount of perceived difference to their self.

We consider changes to intelligence in particular, as in a different experiment (not reported here) we found that changes to intelligence produce large estimations of changes to the self. While 'moral qualities' are reported as important to self image (Strohminger & Nichols, 2014), self-selected changes to them may also carry a moral weight which we do not pursue in our current analysis.

3.6.1 Methods

We recruited 120 participants on Amazon's Mechanical Turk Service (69 male, 37 female, median age = 31, ranging from 21 to 59 years old, 11 participants were dropped from analysis for misunderstanding a question).

Participants were asked to imagine a device that could change their intelligence, making them more or less intelligent by several degrees. However, the device would also change their happiness, such that becoming more intelligent would make them less happy, and vice-versa (the effects ranged from much less intelligent / much more happy to much more intelligent / much less happy, including a 'no change' effect). Participants were asked to indicate their own intelligence level on a 10 point scale (5 was population average), and then choose the effect they desired. The exact phrasing of the question was:

"IMAGINE that there was a Device that works as follows:The Device can make you more intelligent, or the device can make you less intelligent. The Device has magnitudes (a little, a medium amount, a lot). If you become more intelligent, you will also become less happy. The more intelligent you become, the less happy you will be. If you become less intelligent, you will also become more happy. The less intelligent you become, the more happy you will be. QUESTION: How do you use the Device? Do you choose to become..."

Participants were then asked to explain their reasoning (*"Why did you choose the way you did?"*)

On a new form, participants were asked to indicate how different to themselves they would be for different levels of change to their intelligence (regardless of happiness). The intelligence changes were as before, and for each change participants indicated the expected change using a slider ranging from 0 (completely the same) to 100 (completely different). The exact phrasing was:

"Regardless of happiness, how different do you think you would be from your current self, if you were more or less intelligent? For each level of intelligence change, please select the level of difference to your current self. The answers go from 0 (exactly the same) to 100 (completely different)."

Finally, participants provided their age and gender.

3.6.2 Results

Participants generally rated themselves as more intelligent than average (78%, median intelligence rating of 7). Participants saw changes to intelligence as changes to their self, and a larger increase or decrease in intelligence corresponded with a greater change to self. As shown in Fig. 6, participants rated a negative change (decreasing intelligence) as a greater change to the self than a positive change (increasing intelligence).

Regarding their decision to change, 50.5% of participants preferred to change nothing about their intelligence and happiness (N=54), 28% preferred to decrease their intelligence and become happier (N=30), and 21.5% preferred to increase their intelligence and become less happy (N=23). Participants' assessment of their own intelligence was not related to their preference to increase or decrease their intelligence.

Participants' estimation of the degree of change they would undergo was related to the direction in which they would change. Considering participants who would prefer to change, those who prefer to become less intelligent (and more happy) see decreasing intelligence as less of a self-change than those who prefer to become more intelligent (and less happy).

3.6.3 Discussion

On the whole, a majority of participants preferred to remain 'as they were', changing neither intelligence nor happiness. Such a change is not predicted by a naive model of choice, according to which the choice that brings greater happiness should be preferred [9]. We also find that participants in general see a negative change to their self as a greater change. This finding is in line with (Strohminger, Knobe, & Newman, in press), according to which changes for the better are seen as more in line with one's "true self". However,

[9] A possible objection here is that people are distinguishing 'true' happiness from 'fake' happiness. However, to make sense of true and false happiness already requires going beyond a naive model, and we would suggest such a distinction is exactly captured by the machinery discussed in Equation 9. That is, the current self imagines itself as a new, less intelligent, self and finds such a situation abhorrent, even though the self that find this abhorrent would no longer exist if it were actually to undergo the transformation. 'True' happiness in this sense is evaluated relative to the current deciding self.

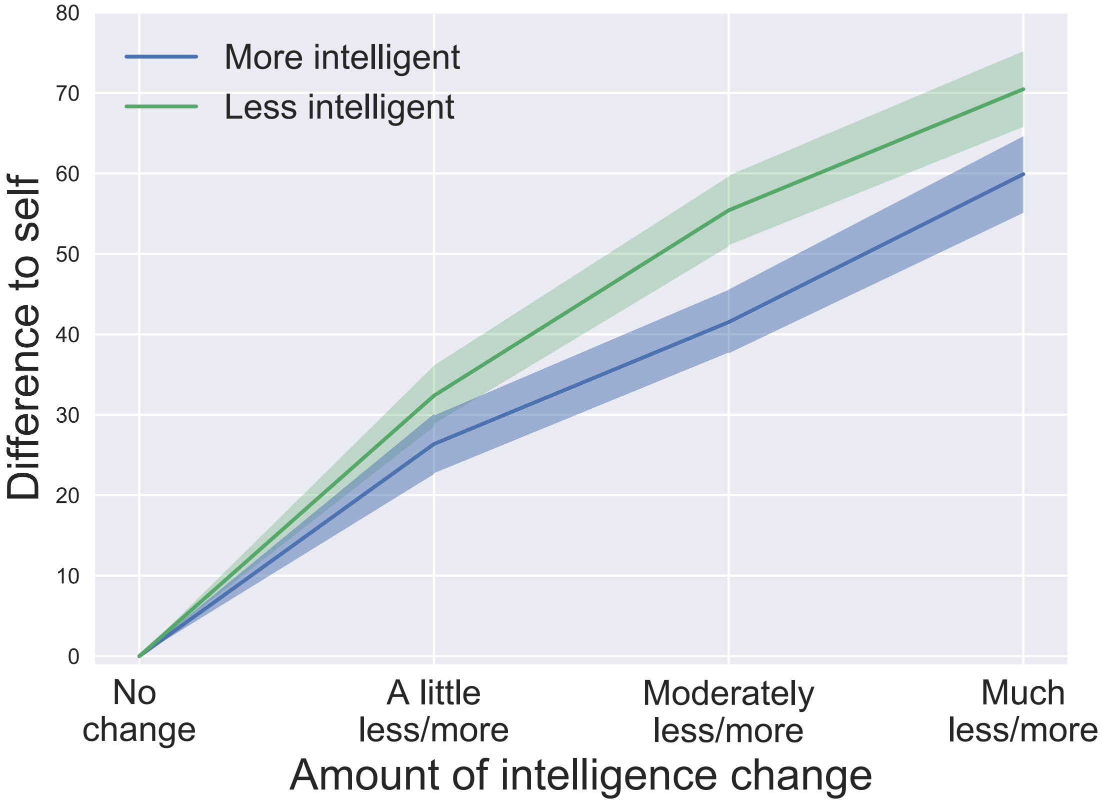

Figure 6: Difference to self ratings for different amounts of intelligence change (either increasing or decreasing). Shaded areas show 95% confidence intervals. Difference increases with amount of change, and becoming less intelligent is seen as a bigger change than becoming more intelligent.

a majority of participants did not wish to change to be more in line with their true self, when it came as a cost of happiness, and a change to their actual self. Again this is partly in line with (Strohminger et al., in press), who argue that expected changes towards the true self are best seen as a gradual process over time.

We found a relation between the decision to change one's intelligence and happiness in a particular direction, and the perceived difference to the self due to intelligence change. More specifically, participants who would change their intelligence for the worse (to become happier) also see negative intelligence change as less disruptive to their self, compared to those who would change their intelligence for the better. This suggests that beyond simple gained utility, the difference to one's self may also play a rule.

Participants' comments were further illuminating. Those who chose not to change expressed a satisfaction with their current intelligence and happiness, including comments such as "I'm quite content with my current self" and "I am the way I am for a reason". Those who would decrease their intelligence focused on the value of happiness for them, with comments such as "I value happiness far more", "Happy is important", and "Happiness is important to me." This suggests people may be treating happiness as a value in its own right, to which a utility can be assigned, rather than being a direct measure of utility. Those who preferred to increase their intelligence focused in part on the increase in opportunities and affordances, with comments such as "Being more intelligent affords me more opportunities", "I'll have more opportunities", "I could achieve more goals than I am now", "Intelligence can take me further than happiness", and "Intelligence has the capacity to help people". Thus, intelligence was seen mainly as an enabling quality rather than a value in and of itself, although one person did explain that "I care about knowledge."

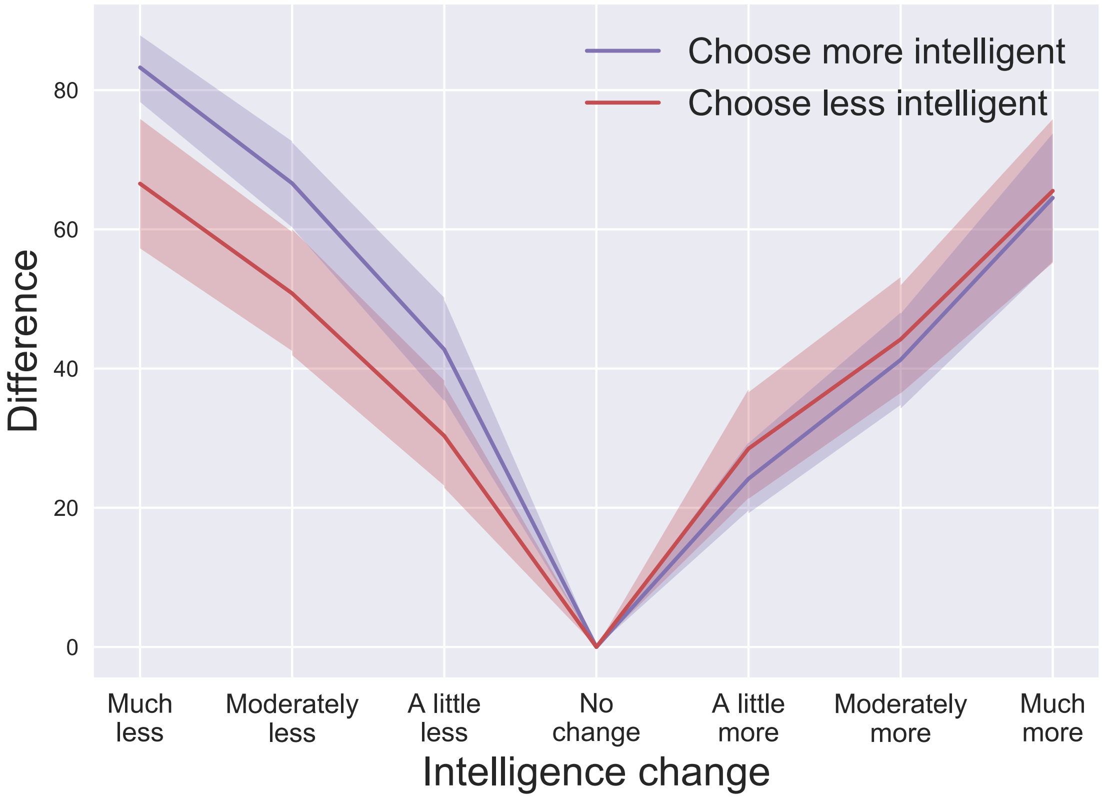

Figure 7: Participants' ratings in Experiment 4 of difference to their self for varying amounts of changes to their intelligence, split by whether a participant chose to become more intelligent, or less intelligent. People who choose to become less intelligent see decreases in intelligence as less of a change to their self.

4 General Discussion

L.A. Paul's recent philosophical inquiry (L. A. Paul, 2014) presents two challenges to any formal account of how people can choose to undergo transformative experiences. The first challenge is that of novelty: A rational approach to decision making requires us to imagine what it would be like to undergo an particular experience, but some experiences (both big and small) are outside our capability to do so. The second challenge is that of change: by transforming we become someone else, with potentially different views on whether we should have undergone that change. The two challenges are related but independent. Even if one knows completely what a new self will be like, there is still a rational difficulty in choosing to become (or not become) that self. In this chapter, we tried to meet both challenges using computational frameworks, which try to ground the everyday intuition that people have when thinking about change and new experience. Our models do not directly address how people ultimately make their decision, but rather how they conceptualize transformative decisions. This division is similar to the one that exists between any intuitive theory and the real world. While we use a theory-of-mind to predict and explain other people and their actions, their actual action-execution and decision-making process may work in a completely different way. Similarly, the way that we explain and predict our own actions may be unrelated to the way we actually make decisions (Saxe, 2009).

Our answer to the challenge of novelty was a formal account demonstrating how people could reasonably choose to experience new things, by leveraging their own preexisting structured view of the world. The model does not, and is not meant to solve the problem of decisions about novel experiences. Rather, the model shows why some decisions about novel experience might be harder than others. When considering an experience that is a sub-category of a more general and well-understood structure (a yellow grape is a grape, which is a food, and so on), people can rely on their past experience, preference, and understanding. For new categories that are higher up in the hierarchy (such as trying a new type of food altogether) people may rely on their hyper-preferences over categories or novel experiences more generally. For decisions that are far outside the realm of understood experience, such as trading a current sense for a new one, our model would be hard pressed to find relevant preferences and experiences to draw on, and it is telling that these decisions also feel intuitively more challenging. Even in such cases, it is also potentially possible for people to leverage their past experiences with preferences and utility change ("I have experienced preference change in the past and things turned out well, so I should try it again")[10]. Such leveraging is not captured by our model, though as with our model is would rely on identifying higher-order features of an experience, which our model can target for having a utility function or preference over.

On top of this, our work highlights the non-trivial nature of even a simple novel decisions, requiring tools from current non-parametric statistics. In this sense philosophy and computational modelling share the purpose of showing the obvious to be non-obvious. People may think it trivial to reasonably choose to eat a new type of grape, but both a philosophical inquiry and a modelling attempt show it is not so. Such an obliviousness to the complexity of the unconscious computations supporting thought is not unique to this particular domain (consider how obvious it seems to hear and understand a friend's words, and then consider the computation that must go into that).

In response to the challenge of change, we built a framework of arbitration between possible selves. While much work needs to be done to further validate these models, empirical evidence does suggest that people do not adopt a strict "view-from-nowhen" when arbitrating between possible selves, and may treat future beliefs and utilities differently. This contrasts with any formalism of transformative decisions that rely upon "global" or perspective-independent decision functions (Pettigrew, 2015) precisely insofar as agents are unable to treat those non-present beliefs and utilities as their own (Moran, 2001).

The aim of our framework in the second half was to capture some of the flavor of thinking about transformative decision-making. In particular, our models address the tyranny of the current self over the potential future self. The tyranny of the present may extend into the past as well as the future: Once a decision has been made to transform, the experiences of the past self are evaluated through the lens of the utilities and beliefs of the current, transformed self. Future work could explore models of agent's post-hoc ratiocination

[10] We thank John Schwenkler for this point.

about transformative experiences and the ways in which it is comparable to ratiocination planning about transformative experiences in light of this tyranny of the present; our present selves did not have time for it.

While we hope our work is useful formally and empirically, much of our work is both suggestive and tentative. In particular, further work needs to be done to disentangle our claims about the reasoning of transformations from the means of those transformations as well as how an agent can verify that a transformation did indeed take place. Our models also leave out several key features unique to big decisions, and it is worth considering the outlines of a formal framework that could in principle address those.

One such key feature is the long shadow cast by the choice not taken (Ullmann-Margalit, 2006). Unlike a transformative *experience*, which may or may not evoke feelings of regret, transformative *decisions* are riddled with counter-factual worries. Our modeling does not specifically address such potential regrets, but in that they are no different from standard utility theories of rational choice under uncertainty, or from the more psychology-based explanations of prospect theory (Kahneman, 2011). There are several proposals for formal models of decision-making that take regret into account (e.g. Bell, 1982; Loomes & Sugden, 1982), but while the causal role of regret is recognized as powerful and important, no current account is as accepted as utility theory or prospect theory. Once such accounts are considered, we may ultimately find that *transformative* decision making does not need any additional conceptual parts to address regret.

Another key feature of transformative decision making is this: Big decisions are hard to do. They are agonizing. People will go out of their way to avoid thinking about them at all, putting off the decision or whittling it down into manageable pieces (Ullmann-Margalit, 2006). When people do bring themselves to think about such decisions, they may cycle through the possible futures, end in indecision, and then repeat the process later. The meta-agents in our framework encounter none of these difficulties or cycling behavior, and it is worth considering in general terms what formal model *could* account for this feature.

A computation may be difficult due to basic limitations of memory and time, such as when a search algorithm must traverse a rapidly fanning tree of options. Thus it may be that, due to the many options involved in imagining two different lives, the subjective analog of this computational difficulty is agony and frustration. But many decisions can be resource-draining in the sense that they involve many options, without being agonizing in the way transformative decisions are. Furthermore, a computationally difficult problem may mean we do want to spend more time and resources getting the answer right, while people actively *avoid* spend time thinking about big decisions. Finally, spending additional time re-visiting a big decision does not seem to produce new imaginings of the future, we learn nothing more about our future selves, but rather go through a cycle of considering the same pros and cons over and over (although there is room for empirically testing this statement).

It may also be that there is no one unique reason accounting for the difficulty of big decisions. People in general dislike making decisions, experience a cost in cost-benefit analysis, shirk responsibility, and deploy second-order strategies to avoid deploying their decision-making apparatus Bobadilla-Suarez, Sunstein, and Sharot (2017); Sunstein and Ullmann-Margalit (1999). If this is the case, models of transformative experience would only have to adopt the amalgam of formal psychological models for the different aspects that make any decision difficult, and inflate them as necessary in terms of cost, responsibility, and the like.

Still, it seems to us that there is one important and distinct feature of the difficulty of transformative decisions, which may require a separate computational analog: Choosing to transform or not to transform forcefully severs the other entertained self. It may be that the state of having future options is in and of itself pleasurable, and cutting off a major branch of a potential time-line is thus painful to contemplate and carry out in and of itself. Our imagined transformed self is not a completely alien being, but rather it is a part of how we see ourselves *right now.* Transformative decisions may be difficult insofar as they force us to let go of that part of ourselves. *Torschlusspanik* is painful in and of itself, but it may particularly painful when we are the ones shutting the gate.

5 Technical Appendix

For the technical appendix, please see: https:www.samuelzimmerman.com/transformativeAppendix.

References

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

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (n.d.). Bayesian theory of mind: Modeling joint belief-desire attribution..

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, C. L., & Tenenbaum, J. B. (2014). Modeling Human Plan Recognition using Bayesian Theory of Mind. In G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, & H. Bui (Eds.), *Plan, activity, and intent recognition.* Morgan Kaufmann. Retrieved from http://web.mit.edu/clbaker/www/papers/baker{_}chapter2014.pdf

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations research*, *30*(5), 961–981.

Bentham, J. (1996). *The collected works of jeremy bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

Bobadilla-Suarez, S., Sunstein, C. R., & Sharot, T. (2017, Jul 27). The intrinsic value of choice: The propensity to under-delegate in the face of potential gains and losses. *Journal of Risk and Uncertainty*. Retrieved from https://doi.org/10.1007/s11166-017-9259-x doi: 10.1007/s11166-017-9259-x

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Gershman, S. J., Malmaud, J., & Tenenbaum, J. B. (2017). Structured representations of utility in combinatorial domains. *Decision*, *4*(2), 67.

Gopnik, A. (1993). Theories and illusions. *Behavioral and Brain sciences*, *16*(01), 90–100.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The cambridge handbook of computational psychology.* Cambridge University Press.

Hamlin, K. J. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, *22*, 186–193. doi: 10.1177/0963721412470687

Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences*, *1001*(1), 134–144.

Jackson, F. (1986). What mary didn't know. *The Journal of Philosophy*, *83*(5), 291–295.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *xx*, 1–16.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1), 99–134.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

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

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

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical {B}ayesian models. *Developmental Science*, *10*(3), 307–321.

Lewis, D. (1990). What experience teaches. In W. Lycan (Ed.), *Mind and cognition: A reader* (pp. 499–519). Oxford: Blackwell.

Locke, J. (1700). *An essay concerning human understanding*. Awnsham and John Churchil, at the Black-Swan in Pater-Noster-Row, and Samuel Manship, at the Ship in Cornhill, near the Royal-Exchange.

Loewenstein, G., & Elster, J. (1992). Utility from memory and anticipation. *Choice over time*, 213–234.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty.

The economic journal, 92(368), 805–824.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. Journal of personality, 60(2), 175–215.

Moran, R. (2001). Authority and estrangement: An essay on self-knowledge. Princeton University Press.

Nagel, T. (1989). The view from nowhere. oxford university press.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. Personality and Social Psychology Bulletin, 40(2), 203–216.

Parfit, D. (1984). Reasons and persons. OUP Oxford.

Paul, L. (2015). Transformative experience: Replies to pettigrew, barnes and campbell. Philosophy and Phenomenological Research, 91(3), 794–813.

Paul, L. A. (2014). Transformative experience. Oxford University Press.

Paul, L. A. (2015). What you can't expect when you're expecting. Res Philosophica, 92(2), 149–170.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology, 5(2), 125–137.

Pettigrew, R. (2015). Transformative experience and decision theory. Philosophy and Phenomenological Research, 91(3), 766–774.

Robinson, T. (1987). Heraclitus: Fragments. a text and translation with a commentary by tm robinson. Toronto: University of Toronto Press.

Russell, S., & Norvig, P. (1995). A modern approach. Citeseer.

Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. The handbook of imagination and mental simulation, 257–266.

Shoemaker, S. (1995). Moore's paradox and self-knowledge. Philosophical Studies, 77(2), 211–228.

Strohminger, N., Knobe, J., & Newman, G. (in press). The true self: A psychological concept distinct from the self.

Strohminger, N., & Nichols, S. (2014). The essential moral self. Cognition, 131(1), 159–171.

Strohminger N., G., Newman, & Knobe, J. (in press). The True Self: A psychological concept distinct from the self. Perspectives on Psychological Science.

Sunstein, C. R., & Ullmann-Margalit, E. (1999). Second-order decisions. Ethics, 110(1), 5–31.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. Science (New York, N.Y.), 331, 1279–1285. doi: 10.1126/science.1192788

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press.

Todorov, E. (2004). Optimality principles in sensorimotor control. Nature neuroscience, 7(9), 907–915.

Ullmann-Margalit, E. (2006). Big decisions: opting, converting, drifting. Royal Institute of Philosophy Supplement, 58, 157–172.

Weirich, P. (2004). Realistic decision theory: Rules for nonideal agents in nonideal circumstances. Oxford University Press.