A computational investigation of being in the world

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Nisheeth Srivastava

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor Of Philosophy

September, 2012

A computational investigation of being in the world

by Nisheeth Srivastava

ABSTRACT

Is there a rational explanation for human behavior? Or is it fundamentally idiosyncratic and beyond the ability of science to accurately predict? In everyday life, we are able to predict the preferences of other people relatively well, and function in a society that is strongly predicated on our ability to do so. Theoretical efforts at predicting how people form preferences, however, have met with repeated failures, resulting in a widespread pessimism regarding the possibility of a universal rational explanation for human behavior.

In this thesis, we provide precisely such an explanation. We show that the errors plaguing existing systems of preference representations are a direct result of the mystery surrounding the actual act of *formation* of preferences, and that once this latter mechanism is clarified, a very large number of paradoxical and contradictory empirical results from the behavioral economics literature are theoretically reconciled. Our investigations lead us to believe that a combination of two simple natural principles is sufficient to both predict and explain why humans make the choices they do: one, that humans seek to always learn what to do in the most statistically efficient manner possible, and two, that this quest for understanding is constrained in remarkably systematic ways by a competing search for choices that can be made with minimal cognitive effort. We find, therefore, that the rational goal that best describes human choice behavior is attempting to minimize the cognitive effort required to make a decision. In other words, in this dissertation, we propose a theory that rational human action is governed by a universal explanatory principle, one that does not match traditional expectations of utility maximization - the principle of least cognitive effort.

This redefinition of rationality has far-reaching implications. In order to better understand them, we constructed an information-theoretic description of a meta-cognitive agent engaging with its environment which allowed us to formulate computationally

tractable intrinsically motivated agents. Simulation studies confirmed that the behavior of cognitively efficient agents provides a possible unified explanation for a large number of behavioral biases identified by behavioral and experimental economists in human subjects, as well as a number of variations in subjects' perception of risk observed in neuroeconomics studies.We further show using mathematical arguments, that this construction is in consonance with existing reinforcement learning literature and, in fact, subsumes multiple strands of current research in broadening the definitions of reward in reinforcement learning. Finally, we extend our analysis to studying social behavior among populations of agents and shed new light on paradoxes in game theory and theories of social interaction, resulting in a demonstration of an amoral basis for being good - the existence of an entirely self-interested (and non-evolutionary) basis for cooperative and altruistic behavior.

In short, this thesis proposes a quantifiable description of agents *being in the world* - detailing universal principles that explain how and why beings develop preferences of the form they do given the structure of the world they inhabit. Our results provide a unification of explanations for several biological and behavioral phenomena spanning neuroeconomics, neurobiology, cognitive science, artificial intelligence and metaphysics.

Acknowledgements

Naturally, the protagonist of this story is my advisor, Paul Schrater. Thank you, Paul, for *being water*. I can offer no greater praise.

The deuteragonist of this story is my uncle, Pramod Srivastava. Thank you, Chacha, for *being water*. I can offer no greater praise.

The tritagonist of this story is a melange of multiple women, who shall, in the interests of both privacy and brevity, remain anonymous. This dissertation investigates the nature of suffering, and so is indebted to this melange, for prolific access to data.

Prof Jaideep Srivastava took me into his lab at a time when I was contemplating dropping out of the program and lavished me with undeserved and unstinting support for three long years. Though I regret not having been more productive for him in exchange for his unfathomable generosity, he has my gratitude in perpetuity.

I am grateful to Arindam Banerjee for bringing me to Minnesota. It is clear to me that, even with the benefit of hindsight, I could not have designed more propitious conditions for my work. While eventually our philosophical differences proved insuperable (he believes that graduate students should work, I believe that PhD stands for 'paid holiday'), I enjoyed our brief engagement, and the lossy compression ideas that we started working on in my first year here have made their way, subtly, into the structure of the theory that I have proposed in my dissertation work.

I am grateful to the CS program at the University of Minnesota, for holding my feet to the fire when necessary, and forcing me to do creative things fast to avoid getting kicked out. I am grateful to the custodial staffs in Keller and Elliott Hall, for various favors that are best left undescribed and to the University of Minnesota campus, for being *home* to me for five years, in every sense of the term.

Last, but not least, I am grateful to the Mississippi, for being there.

Dedication

To Shreya,

For breaking me

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consider a young child tossing a ball up in the air and catching it on its way down. As he gets older, he will learn how to catch balls in his vicinity in increasingly more complex situations: when playing catch, when playing baseball, etc. How does he manage this? There is a regularity in the way the ball goes up and then comes down. Over repeated tosses, the child will learn where the ball is likely to be at a particular point in time beforehand. That is, he will *predict* its future location. Then, he will figure out a way to move his arm, hand and fingers to grab it.

Consider now the case of a young child trying to wheedle his mother into letting him eat one more cookie. The mother, naturally fearing his sugar high, is originally disinclined to give him one. The child wheedles. He promises to be good; he promises to wash the dishes, etc. When bargaining doesn't work, he uses, successively, coaxing, pleading and finally, emotional blackmail, "You don't love me!". In our story, the child ultimately succeeds in wearing down his mother's defenses, and procures the additional cookie.

How did he manage this feat? Over repeated encounters with his mother, he developed a sophisticated theory of negotiation. Most mothers will attest that some of their children know exactly how to get them to do what they want. In effect, the child in our example could *predict* his mother's future behavior, and adaptively planned his negotiating strategy bearing those predictions in mind.

The common thread between these two scenarios is the ability of the child to predict the behavior of his environment and adjust his actions accordingly. This is, in a nutshell,

the definition of *intelligent* behavior. The study of intelligence is the study of the ability of natural and/or artificial agents to learn about regularities that exist in the world, and exploit them in ways that promote their goals.

Beings differentiate between possibilities in the world by selecting to participate in some and not others. Intelligent behavior, to the extent that it can be specified without reference to unobservable mental properties, is simply choosing possibilities in the world wisely. Beings intentionally select some possibilities and not others because they believe that doing so is useful. This dissertation identifies universal natural principles that explain how such *preferences* change, as the world outside the being changes.

As we illustrate in Figure 1.1, the intrinsic human ability to predict physical dynamics has been subsequently systematized and mathematized to yield simple and profoundly beautiful descriptions of the nature of the physical universe. A natural question, very important, but rarely contemplated, is: **why** does the universe, to an almost insane degree of precision, appear to follow such mathematically simple physical laws? There are two logical categories of answers to this question: one that says that there is no real reason, another that says that there probably is one. In the first case, all further inquiry must cease. It is the second case that holds out promise for possible intellectual inquiry. Parsimony suggests that the metaphysically lawful behavior of the physical universe must also extend to the behavior of carbon-based phenotypically adaptive organisms within it. Whether the fundamental explainability of the universe has an explanation or not is beyond the scope of this work. However, if there is one, we expect it to hold across all natural phenomena. Thus, metaphysical parsimony suggests that there are *universal principles* that explain human behavior. We just haven't found them yet.

Such was the faith that motivated the work that is summarized in this dissertation: that human behavior is not irrational, and therefore must be predictable in the same way as physical variables, as arising out of some universal natural principles. Unfortunately, as we also describe in Figure 1.1, the history of the study of behavior has not led to theoretical discoveries of the kind seen in the physical sciences. At present, the idea that human behavior is rational, as expressed in the *homo economicus* tradition, is no longer seriously entertained in the research community.

The roots of this failure lie deep in the past, at the very foundations of post-Enlightenment critical thinking about behavior. From the outset, theorists carved out

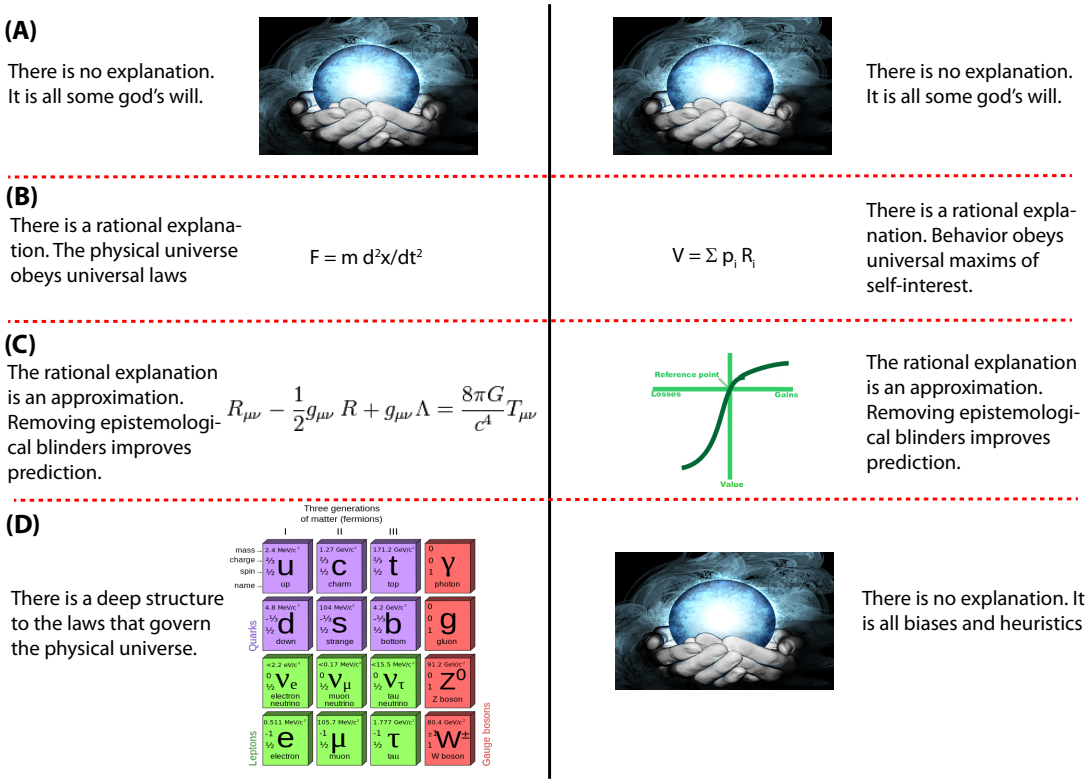


Figure 1.1: Similarities and differences in the respective histories of the physical and behavioral sciences. **(A)** Pre-rational societies believed that the physical world was determined by some wilful agency, and hence, could not be predicted. In the same way, nothing could be predicted about free human behavior, except that it would annoy the wilful agent in some way. **(B)** The Enlightenment revolutionized the intellectual aspirations of humanity, demonstrating through the work of luminaries such as Pascal, Newton and Leibniz that mathematics could actually explain many disparate phenomena in the world. Following in this tradition, moral philosophers like Mill and Locke asserted that similar universal priniciples might also govern human behavior. **(C)** The original physical theories of the Enlightenment era ignored the mechanistic underpinnings of the variables that they were trying to describe. Beginning in the first decade of the 20^{th} century, a series of insights into the transmission mechanisms underlying various field effects transformed the world of physics. Similarly, the epistemological premises of behavioral theories of rationality were also questioned, and several alternative descriptions of behavior were proposed. **(D)** Unfortunately, whereas the powerful insights about the epistemology of the physical world have culminated in a deep, accurate and fulfilling understanding of the physical universe, behavioral theory appears farther from rational specification than ever before, with multiple experiments showing that subjects' behaviors cannot be constrained within any known theoretical framework.

a distinction between the *acquisition* of human preferences and their adequate *representation*. The former problem was considered to be too hard for meaningful study; investigations of the second problem led to the foundations of economic theory in the late 19^{th} century, as well as mathematical psychology in the mid-20^th century. However, even as these foundations were being settled, theoretical and empirical critics of both programs began denouncing their rationales as flimsy, demonstrating violations of behavior expected under classical economic and psychological predictions. While early critiques were considered accounted for by the intrinsic variability of behavior and the lack of granularity in experiment designs, seminal research by behavioral economists beginning in the 1970s conclusively established that human behavior was, in fact, systematically biased against the predictions of existing theories. Such findings should have caused behavior theorists to revisit their assumptions about the foundations of human behavior. Instead, what resulted was a schism between *methodologists* who continue, to this day, to use the mathematical structures of classically rational preference theories to build predictive models (with parameter fitting serving to reduce disparities between theory and prediction), and *anthropologists*, who turned away from the hope of finding universally rational descriptions of human behavior towards research that attempts to document heuristics that appear to predict behaviors in various decision scenarios, in the hope that once a comprehensive collection of such observations has been procured, we will eventually be able to pick out which heuristic will best explain a given behavior pattern out of a comprehensive list [1]. Such theories, while often more accurate than rational models, give up on the possibility of universal explanations for human behavior. Thus, we find that the current state of research in theories of preference is best described as a world where the technicians build ever more sophisticated models detached from reality, while theorists resort to stamp-collecting[1] .

This dissertation has attempted to seek out a principled foundation for explaining human behavior, beginning from the precise point in the history of the process where we believe existing theory fell into error, viz., the beginning. We believe that the behavioral sciences never underwent the natural philosophical revolution that physics did, wherein questions about 'how' theories worked proved instrumental in addressing misconceptions

[1] This references Ernest Rutherford's famous observation, "All science is either physics or stamp-collecting', and references stamp collecting precisely as he meant it in his original usage.'

about the variables that physical theories were trying to predict[2] .

We believe that, having postulated that human behavior should maximize self-interest (called utility) in order to be rational [2], behavioral theorists failed to account for the epistemological aparatus that humans possess in constructing their theories of behavior. Thus, trivial instantiations of self-interest maximization (e.g. maximize money) were passed off as approximations of theories of human behavior. While subsequent research has clearly shown such theories to be faulty, the epistemic assumptions underlying them have seldom been questioned. The origins of theoretical attempts to understand human behavior as being governed by concepts like value, utility, probability etc. date back to the seminal mathematical treatises of Bernoulli. Ever since, theorists have been trying to specify what value, utility etc. mean in ever more mystifying conglomerations of mathematical assumptions.

The more recent realization of the futility of trying to describe behavior in terms of utility maximization has led theorists to propose that human preferences arise out of experience with the environment, and that choice strategies that are ecologically suitable for a particular set of choices that an agent (or an evolutionary ancestor) has made in the past may appear irrational from a utility maximizing perspective in other choice contexts, explaining the violations of theoretical expectations seen by behavioral economists. Thus, there have been several recent efforts to elicit rational explanations for various non-normative behavior patterns based by taking typical environment statistics into account. A prototypical example of such an exercise is Stewart, Chater & Oaksford's 'decision by sampling' framework [3], where they show how, for instance, typical wealth distributions in the population can lead endogenously to risk-averse agent preferences, and typical statistics of stock market returns can explain asymmetric gain vs loss aversion, as documented in the prospect theory literature [4]. While such work certainly advances in the right direction of seeking rational explanations for behavioral phenomena, in practice, all the explanatory power in its explanations lies in the statistics of the environment, with the agent exerting no real *choice* in the matter. Thus, while

[2] For instance, Einstein's deep insight into the relativity of reference frames was sparked by trying to understand what an observer perched on a photon would be able to see. Similarly, trying to understand how fields could, seemingly magically, affect particles across space and time, led to the development of quantum electrodynamics, the crown jewel of physics and direct precursor to quantum chromodynamics and the astoundingly pretty Standard Model.

such work accentuates the importance of taking environment properties into account while addressing preference formation, it does not make any interesting endogenous predictions, and so, is judged insufficient.

A more mathematically sophisticated theory of preference formation emerges from the work of Karl Friston [5], who argues that human behavior, at multiple levels of biological organization, is fundamentally extropic, i.e. attempting to minimize entropy. Fristonian agents operate on the simple guiding principle of trying to minimize suprise, which leads to a surprisingly large number of useful and interesting predictions. However, in its current form, it necessarily predicts that agents will always avoid surprising situations, which contradicts both plain common sense and a large literature consistently demonstrating curiosity, novelty-seeking etc. in humans. Therefore, while we are personally extremely symapthetic to his work, and believe that our own theory is quite likely deeply related to his proposal, in practice, it has not yet been studied in economics and behavioral contexts, and so, cannot be evaluated completely.

A common thread runs through various efforts to reconcile agent behavior with theoretical predictions - increasingly sophisticated coupling with the environment. Thus, while plain utility maximization assumes indifference towards the sources of preferences, the adaptive toolbox approach to choice modeling acknowledges their importance, but remains agnostic about the exact mechanisms by which these sources inform the agent. Decision by sampling approaches further specify that agent behavior is governed by the statistical properties of the environment, i.e. which events occur most often, but use simple mathematical techniques resulting in the prediction of probability matching under all circumstances, which is unrealistic. The latest breed of behavioral theories assume much more sophisticated relationships between the environment and the agent. Friston's theory, as we have noted above, argues that agents try to minimize the entropy of the event distribution they experience. Gershman & Daw [6] arrive at the Fristonian objective function (free energy) using other assumptions about the underlying theoretical mechanism.

Our own research, which chronologically precedes [7] this latest generation of preference theories, introduces a hitherto underappreciated notion idea to this trajectory of increasing agent-environment-coupling sophistication. We allow that agents **learn** what to do from their environment. While such a statement might appear surprising

at first sight, it is a fact that no previous theory of preference formation has allowed agents to learn what to do from the environment! Learning-based models of behavior like reinforcement learning [8], experience-weighted attraction, [9] etc. all assume that agents can directly estimate *rewards* embedded in the environment. By doing so, they immediately hide the problem of preference acquisition within the reward-inference machinery of the agent. Once we imbue an agent with the ability to assign numbers to possibilities, then being able to figure out how to extract out the biggest number, even in the presence of uncertainty and multiple trials, is just an exercise in statistical estimation, not a theory of how humans might actually behave. In light of the common superstition that artificially intelligent planners and learners are somehow actually 'intelligent' in some meaningful way, it is important for the Computer Science reader to understand this distinction at the very outset of this dissertation. AI agents are about as intelligent as railway trains. If you lay out tracks for them to reach their destination, they will arrive safely. We note that this criticism is not original to this dissertation. Hubert Dreyfus' trenchant critique of the limitations of GOFAI largely anticipate our comments by about three decades [10].

To return to our original point, an original contribution of this dissertation is the blending of machine learning into theories of preference formation. While it would be hard to find a sane person who would dispute that human preferences **are** indeed, learned from experience to a considerable extent, the mathematics for how such learning would proceed in practice has remained unknown. We fill in this gap by developing agents that learn what to do in the world, based on their past experience.

A knowledgeable reader would, at this point, pause in puzzlement. "This sounds too simple. Why has it not been done before?" Recall the difficulty that [1] and [3] have had in trying to couple even highly specific tasks to the information that agents should extract from them. Friston [11] succeeds in doing this for any possible environment by assuming that the only relevant information that agents need to extract from the world is the statistical frequency of possibilities. Clearly, the former approaches cannot be generalized, while the latter approach has been rightfully criticized as being over-general [12]. Ultimately, therefore, the problem with developing the right agent-environment coupling is figuring out what aspects of the environment are substantially universal to admit constructing generalized mathematical abstractions from. Friston

says statistical frequency, and that doesn't work. So what could?

The idea that it is the interface of the agent and the environment that yields insight into behavior has been well-understood in many philosophical traditions dating back to antiquity[3] . Aristotle's gigantic shadow suppressed consideration of this idea in Western intellectual thought for 2000 years, until Martin Heidegger identified it as fundamental to being able to meaningfully discuss phenomenologically salient concepts such as *being* and *existence* [13]. At the inception of this research, we realized that both Heidegger and Buddhist metaphysics offer a succinct and, more importantly quantifiable, description of the universal characteristics of the agent-environment coupling. Remarkably, in my interpretation of both sources, Heidegger appears to have rediscovered precisely the triplet of ontological properties as the Buddhist seers had suggested.

The research described in this dissertation documents our effort to construct a learning framework that parsimoniously accommodates the ontological primitives illustrated in Figure 1.2. Our vision rapidly coalesced around the intuition of a meta-cognitive agent, required by evolutionary selection pressures to form accurate preferences to help it navigate its world, and capable of expressing dissatisfaction with its past preferences about the world. An agent that must always **learn** what to do based on its experiences cannot impute any value or judgment to possibilities in the environment that are not determined through its own relationship with them. Hence, such a learning agent is characterized by *existence*, in the Heideggerian sense[4] . The doctrine of *anatta* may be recognized as an antinomian recognition of the same fact, that there is no objectively independent existence for any mental phenomenon. Because *Dasein*, being-in-the-world, is situated in the world with entitities not predictively co-extensive with it, it will fail to impute accurate preferences for the behavior of those entities, hence being characterized by *fallenness*. The dissatisfaction experienced by this failure (which is inevitable and continuous throughout the agent's existence) characterizes *Dasein* by *dukkha*. Concomitantly, because *Dasein* is not co-extensive with other entities, it will experience their existence as transient, thereby living in a world characterized by *anitya*. Finally,

[3] In fact, assertion of the the unity of the subject-object is perhaps the one universal principle across multiple Eastern mystical traditions, e.g., Yoga, Advaita, Mahayana Buddhism, Zen and Daoism.

[4] The philosophically agnostic reader need not despair. We will make no further reference to metaphysics or philosophy beyond this paragraph until the final chapter of this dissertation. Thus, we emphasize that our scientific and technical contributions stand independent of their metaphysical motivations.

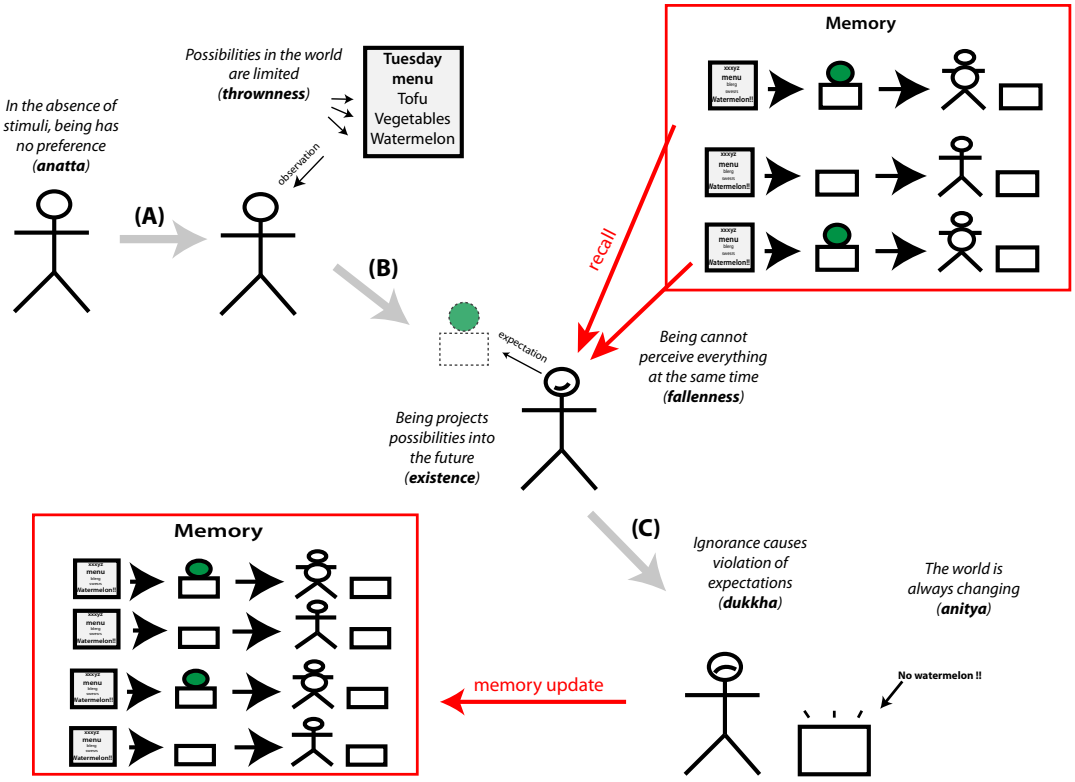


Figure 1.2: A meta-cognitive agent that learns what to do based on past experiences with the world satisfies the universal phenomenological characterizations of *being in the world* separately identified by Martin Heidegger and Buddhist metaphysicians. In this illustration **(A)** an agent holds no preference for watermelons in the absence of the possibility of watermelons being accessible. Once the world reveals the possibility of watermelon availability, the agent **(B)** projects its past experiences of such revelations to form a preference for procuring watermelon. Since the world is dynamic, and the epistemic capabilities of the agent are limited (it hasn't seen all possibilities, and it cannot recall all the ones it has seen) **(C)** it will often develop erroneous preferences. A metacognitive agent will be able to track the performance of its preference-driven choices and experience dissatisfaction when they prove inaccurate.

it is assumed that a realistic agent will never be able to experience all possibilities, its cognitive apparatus will be epistemologically constrained, leading to the characteristic experience of *thrownness*[5] .

It is not sufficient to simply introduce a learning mechanism in a preference elicitation setting to retrieve a rational theory of preference formation. A rational theory requires a rationalizable objective function. In standard utility theory, this was represented by cumulative or average utility, with the agent's rational goal being to maximize it. If we eschew using utility functions, what could replace it? As we detail more extensively in Chapter 4, we discovered an evolutionarily plausible principle of least action to replace utility maximization - the principle of least cognitive effort. Our theory of preference dynamics states that *humans make choices that involve the least cognitive effort without reducing their confidence in the ability to make accurate predictions.* We call agents that obey this principle of rational action *cognitively efficient agents*, and the principle itself the cognitive efficiency principle.

By replacing external utility with an internal measure of cognitive effort, we yoke our agent's objective function to the agent-environment interface, as desired. At the same time, we developed information-theoretic measures of cognitive effort and confidence that allow us to generalize our new principle of rational action across different action domains, which is where previous efforts to extend rational analysis to account for environmental features failed. Along the way, we found that a proper specification of preference representations depends fundamentally on the underlying process of preference acquisition (as we detail in Chapter 2), and that many of the errors of the *homo economicus* research program have arisen principally out of the historical separation of these two aspects of the basic problem of explaining preference dynamics.

Chapter 2 describes what happens when we assume that subjects cannot see all world possibilities at the same time. We show that postulating efficient value inference in this setting leads to an endogenous explanation for all known framing and/or context effects observed in human subjects, as well as a natural and representationally equivalent replacement of utility functions as descriptors of human behavior. Finally, we show also that simply assuming limitations in observing external events together does not account

[5] We titled this dissertation referencing 'being-in-the-world' to honor these motivating principles behind our research, removing the hyphens to indicate that we believe that our work has substantially clarified various obscurantist interpretations of Heidegger's original concept.

for a number of other biases documented in the literature, particularly probabilistic ones. These results have been partially reported previously in [14]

This weakness of simple inductive rationality sets the stage for us to introduce the idea of cognitive limitations that further constrain preference formation. In Chapter 3, we describe how simple inductive inference, of the form proposed in Chapter 2, fails to explain intriguing human data, forcing us to examine further constraints on the epistemological capabilities of agents. We show that adopting a cognitive efficiency criterion for memory recall explains the data, and offers a novel explanation for the origins of risk appetites in humans.

Chapter 4 contains the main exposition of our theory, wherein we pull the pieces of inductive inference and cognitive limitation together to develop a new theory of belief/preference dynamics. We show that our new theory leads to endogenous predictions about behavior replicating major categories of documented cognitive biases in human subjects, as well as a demonstration of the functional identity between decisions from experience and decisions from description using data from a large empirical study [15]. A partial account of these results has been previously reported in [16] and [17].

Chapter 5 draws connections between our theory of cognitively efficient learning and existing reinforcement learning based accounts. We show how our theory of belief dynamics can be further generalized as a form of active inference, closely related to the theory of reinforcement learning. This results in a formal specification of goal-directed machine learning that better describes decisions taken in real-world situations by real organisms.

Up to this point in the dissertation, we consider only single agents that are learning what to do in an environment populated by uninteresting 'entitites'. What happens if the environment contains other agents just as well? In Chapter 6, we answer this question by extending our theory of belief dynamics to develop a theory of social preference formation. Simulated experiments show that our theory of cognitive efficiency shows a way towards eliciting cooperative and altruistic behavior in agents that are indifferent to the outcomes experienced by other agents. Our results suggest that postulating social utilities is unnecessary, and that altruistic and cooperative behaviors need not necessarily have evolutionary antecedents. These results have been partially reported previously in [18].

We conclude with a brief description of possible applications for our theory in Chapter 7.

As the essential contribution of this dissertation, we report our discovery of two simple natural principles that combine to explain a large cross-section of experimental data capturing human behavior in different task settings.

1. **Principle of efficient learning** Humans try to learn what to do in the most statistically efficient way possible.

2. **Principle of cognitive least action.** Humans base their decisions on what to do in ways that minimizes them having to remember past experiences, which constrains the statistical efficiency of their learning in systematic ways.

Other contributions, more specialized to different scholarly communities, will be highlighted in individual chapters. The empirical results reported in this thesis lead us to believe that these principles are potentially deep, and reflect universal physical laws that govern human choices. We therefore believe that the net contribution of this research to the state of current knowledge will be significant.

Throughout our exegesis, we focus entirely on describing behavior in terms of *preferences* that observers acquire about different possibilities (or affordances) they can experience in the world. *Beliefs* refer to mathematical objects that can contain these preferences, and *decisions* refer to physically observable behavior, selecting one amongst many choices presented.

As a brief reading guide, readers interested only in understanding the basic outline of our theory may skip directly to Chapter 4. Readers primarily interested in the economics contributions of our work may focus profitably on Chapters 2,3 and 6. Readers with a psychology background will chiefly enjoy Chapters 2,4 and 6, while computer science specialists will probably find the most value in Chapters 4 ,5 and 6. Game theorists and social scientists will find it best to skip directly to Chapter 6, which is written in a self-contained manner, including descriptions of the individual preference dynamics theory.

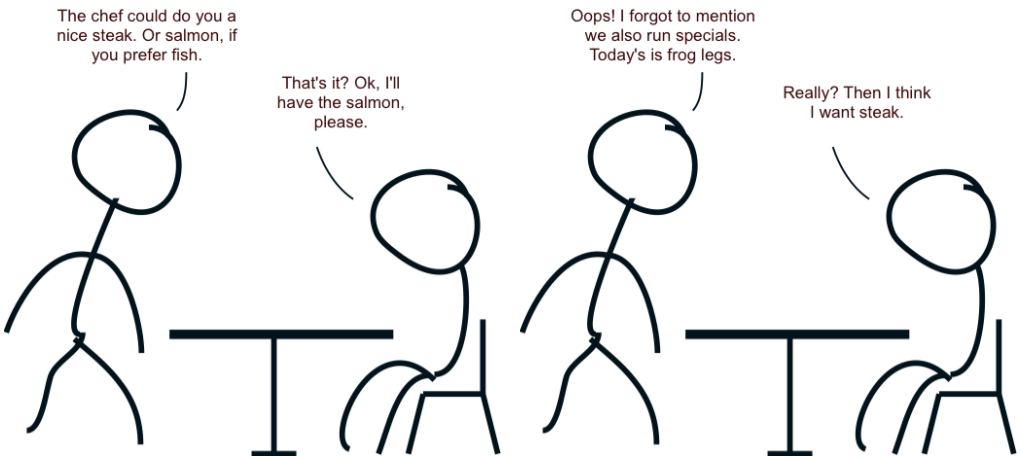
Chapter 2

Rational inference of relative preferences

Normative theories of human choice behavior have long been based on how economic theory has postulated they should be made. The standard version of the theory states that consumers choose *rationally* using innate, stable preferences over the options they consume. Preferences are represented by numerical encoding of value in terms of utilities, and subjects are presumed to select the option with the maximum expected utility. The mathematical simplicity of the expected utility framework has allowed it to maintain a central role in microeconomics [19], machine learning [8], computational cognitive science [20] and neuroscience [21]. So long as preferences could simply be treated as methodological abstractions in theoretical economics, it was considered reasonable to simply assume their existence, and that their revelation via rational subject choices provided sufficient epistemological basis to justify their usage as descriptions of human behavior. However, with advances in the neuroscience of decision-making afforded by new imaging technologies and experimental protocols, researchers are increasingly drawing closer to asking fundamental questions about the actual origins of human preferences. Many of the answers they have been receiving cast foundational doubts about the basic premises of economic theory. In this paper, we try to resolve one such difficulty - developing a representation for the brain's encoding of preferences that reconciles new behavioral and neurobiological evidence with the theoretical expectations of economic

theory.

Human preferences exhibit patterns of behavior that are impossible to reconcile with the idea that stable numerical representations of value can be ascribed to each item they choose between. Evidence for this impossibility comes from both behavioral and neuroscientific experiments, which we here briefly review. Behavioral experiments in the last half century have conclusively demonstrated (see [22] for a comprehensive review) that human choice strongly violates the key axioms that the existence of stable utility values depends on. A particular subset of these violations, called context effects, wound the rational choice program the most deeply, since such violations cannot be explained away as resulting from epistemologically limited access to monotone distortions of underlying utility and/or probability representations, as afforded by various rank-dependent generalized utility theories [22].



(a) When asked to select between just salmon and steak, the diner picks salmon, indicating salmon $\succ$ steak by his choice

(b) When presented with an additional third menu item, the diner picks steak, indicating steak $\succ$ salmon

Figure 2.1: Illustration of Luce's 'frog legs' thought experiment. No possible absolute utility assignation to individual items can account for the choice behavior exhibited by the diner in this experiment. The frog legs example is illustrative of reversals in preference occuring solely through variation in the set of options a subject has to choose from.

Consider for instance, the "frog legs" thought problem, pictured in Figure 2.1, introduced by Luce and Raiffa in their seminal work [23]. No possible algebraic reformulation

of option-specific utility functions can possibly explain preference reversals of the type exhibited in the frog legs example. Similar preference reversals elicited through choice set variation have been observed in multiple empirical studies, using a variety of experimental tasks [24, 25].

Turning to the neuroscientific evidence, reinforcement learning driven accounts of dopaminergic pathways suggest that they carry an error signal for updating a moving average of experienced outcome valuations [26]. The targets of these neurons, notably in the striatum and prefrontal cortex, are believed to be involved in valuation and action selection [27]. However, evidence for the existence of stimulus-specific value coding in the brain is weaker. Vlaev et al [28], reviewing recent neurobiological literature, point out that there is no evidence that the brain possesses a common neural currency to evaluate options independently *across* contexts.

Intriguingly, some studies demonstrate the existence of neuron populations sensitive not to absolute reward values, but to proferred options being better relative to other options, a phenomenon called comparative coding. Comparative coding was first reported in [29], who observed activity in the orbito-frontal neurons of monkeys when offered varying juice rewards presented in pairs within separate trial blocks in patterns that depended only on whether a particular juice is preferred within its trial. Elliott et al. [30] found similar results using fMRI in the medial orbitofrontal cortex of human subjects a brain region known to be involved in value coding. Furthermore, comparative coding has been observed under both rewarding and aversive outcomes [31].

Both behavioral and neurobiological studies also suggest that humans possess two basic modes of stimulus response - gain/pursuit and loss/avoidance. For instance, [32] have shown that humans frame their responses to gambles in terms of *gains* and *losses*, measured from a subject-specific set-point. Subject decisions in these two different frames differ significantly, indicating that different evaluative systems operate in either regime. Gray & McNaughton [33] present evidence showing activity in different brain regions during responses to rewarding and aversive stimuli.

These observations summarily imply that any realistic description of the brain's encoding of value must satisfy at least the following desiderata:

1. It must be able to represent *comparative coding*, viz. choices made by human subjects that appear to be sensitive to comparisons across available choices, not

to magnitude of reward obtained from the selected object.

2. It must be able to represent *utility coding*, viz. it should be able to explain choice behavior that appears to be driven by sensitivity to standard utility functions.

3. It should be able to explain the effect of framing choices as gains or losses, as well as other *context effects*, on subjects' preferences.

The principal contribution of this work is the development of a model that **infers** preferences from limited information about the *relative* value of options. We show that we only have to postulate that feedback from decisions provides limited information about the relative worth of options within the choice set for a decision to retrieve an inductive representation of value that is equivalent to traditional preference relations. Thus, instead of assuming utilities as being present in the environment, we learn an equivalent sense of option desirability from directly observable information in a limited format that depends on the set of options in the choice set. By relying entirely on observable choices as sources of information, our preference elicitation methodology remains **rational** - in the general sense that an agent's choice is enforced upon it, given its history of observed choices. This redefinition of rationality allows us to infer relative preferences that are normative, yet uncircumscribed by the infamous von Neumann-Morgenstern [34] axiomatic constraints. This inductive methodology naturally makes choice sets informative about the value of options, and hence affords simple empirical explanations for context effects observed in human subjects. We also show conditions under which our generalized sense of option desirability becomes representationally equivalent to traditional utility measures and how it can be used as a replacement for expected utility valuations in risky decisions, thereby recovering the basic functionality of standard utility functions. In short, in this work, we show how to formalize the idea of relative value inference in a way that is sensitive to the epistemic limitations of humans and that it provides a new **inductively rational** foundation for understanding the origins of human preferences.

2.1 Human preferences via value inference

Let $p(x|r^{(1:t)})$ represent the belief a value x is the best option given a sequence of value signals $\{r^{(1)}, r^{(2)} \cdots, r^{(t)}\}$. Since the agent learns this distribution from observing $r(x)$ signals from the environment, an update of the form,

$$p(x|r^{(1:t)}) = p(r^{(t)}|x) \times p(x|r^{(1:t-1)}), \tag{2.1}$$

reflects the basic process of belief formation via value signals. When value signals are available for every option, independent of other options, the likelihood term $p(r|x)$ in Equation (2.1) is a probabilistic representation of observed utility at time t, which remains unaffected in the update by the agent's history of sampling past possibilities and hence is invariant to transition probabilities. Such separation between utilities and probabilities in statistical decision theory is called *probabilistic sophistication*, an axiom that underlies almost all existing computational decision models [21]. The probabilistic form $p(r|x)$ of standard utility functions has been previously shown to satisfy the axiomatic requirements of Savage's utility axioms [35] by interpreting r as representing a random variable indicating the satisfaction of some need and x as an action possibility [36].

The crux of our new approach is that we assume that value signals $p(r|x)$ are *not available for every option* . Instead, we assume we get partial information about the value of one or more options within the set of options c available in the decision instance t. In this case value signals are hidden for most options x. However, the set of options $c \in \mathcal{C} \subseteq \mathcal{P}(\mathcal{X})$[1] observed can now potentially be used as auxiliary information to impute values for options whose value has not been observed.

What can we say about the value signal itself? As we mention above, any realistic encoding of the brain's sense of value must be able to account for both comparative and value-based results. An additional consideration arises if we further seek to ensure that our account of preference formation be scientifically falsifiable, viz. that all our inputs remain directly observable. In discrete-choice settings, to which we confine our attention in this paper, the only directly observable measurement we have of rational agents' behavior is which option they considered most desirable. Any falsifiable theory of value inference must be grounded entirely in observable quantities, enforcing a selection

[1] $\mathcal{P}(\cdot)$ references the power set operation throughout this paper.

of the simplest possible form of comparative coding - the ability to identify the best (or worst) of observed options.

Instead of computing absolute utilities on all $x \in \mathcal{X}$, therefore, context-aware agents in our framework evaluates the comparative *desirability* of only those possibilities considered feasible in a particular context c. Appropriate semantics for defining such a desirability *pointer* pre-exists in the form of Kripke frames $\langle C, R \rangle$, where C is a non-empty set, and R is a binary relation on C, typically called an *accessibility* relation. We accomodate gain/loss polarity in value perception by permitting accessibility relations of two types - gain-seeking relations that *point* to the best available option and loss-avoidant relations that *point* to the worst available option. Given a choice set C, the choice of the accessibility relation further frames the desirability function into either a gain or a loss setting for the agent to compute desirabilities. The choice of which accessibility relation to use in a particular context will depend on decision-theoretic machinery outside the purview of our present exegesis, which focuses on value representation. This formulation easily allows representation of comparative coding and context effects. In 2.2.2, we show how this value encoding can also represent utilities with a few additional assumptions, thereby satisfying the desiderata we have previously specified.

Thus, instead of using scalar values to indicate which possibility is more preferable, we introduce preference information into our system via a desirability function d that simply *points* to the best option in a given context, i.e. $d^{(c)} = B$, where B is a binary relation (c, c, m) and $m_i = 1$ iff $c_i \succ c_{i'} \forall c_{i'} \in c \setminus \{c_i\}$ and zero otherwise[2] . The desirability indicated by $d^{(c)}$ can be remapped on to the larger set of options by defining a relative desirability across all possibilities $r(x, c) = m, x \in c$ and zero otherwise.

Recall now that we have already defined utility as $p(r|x)$ in our system in Equation (2.1). Instantiated in the discrete choice setting under partial observability of world possibilities, this can be restated as a probabilistic definition of relative desirability as,

$$\boxed{R(x) = p(r|x) = \frac{\sum_c^{\mathcal{C}} p(r|x,c)p(x|c)p(c)}{\sum_c^{\mathcal{C}} p(x|c)p(c)},} \tag{2.2}$$

where it is understood that the *context* probability $p(c) = p(c|\{o_1, o_2, \cdots, o_{t-1}\})$ is a distribution on the set of all possible contexts inferred from the agent's observation

[2] For aversive stimuli, the accessibility relation would be defined using an $m^{(l)}$ such that $m_i^{(l)} = 1$ iff $c_i \prec c_{i'} \forall c_{i'} \in c \setminus \{c_i\}$ and zero otherwise.

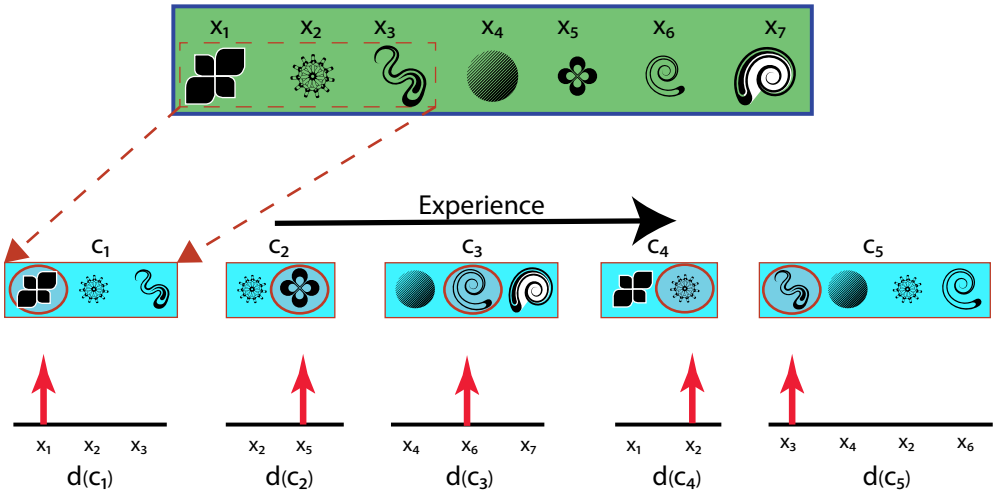


Figure 2.2: Possibilities **(x)** in the world are never observed together. Instead, they tend to co-occur in subsets, indicating particular contexts **(c)** of observation. Our theory of value inference considers information about the desirability of options to be confined to desirability functions **(d)** that point to the most (or least) desirable option in a given context.

history. From the definition of desirability, we can also obtain a simple definition of the *desirability* probability $p(r|x, c)$ as $p(r_i|x_i, c) = 1$ iff $r_i x_i = 1$ and zero otherwise.

Term	Name	Interpretation
$p(r\|x, c)$	Desirability probability	Probability of possibility x being desirable within context c
$p(x\|c)$	Observation probability	Probability of observing possibility x, given observation context c
$p(c)$	Context probability	Prior probability of deploying observation context c

Table 2.1: The relative desirability of an option across all observed contexts can be interpreted as a combination of three probabilistic contributions.

To instantiate equation (2.2) concretely, it is finally necessary to define a specific form for the *observation* probability $p(x|c)$. While multiple mathematical forms can be proposed for this expression, depending on quantitative assumptions about the amount of uncertainty intrinsic to the observation, the underlying intuition must remain one that obtains the highest possible value for $c = o$, o being the subset of possibilities that is actually observed at the decision instance, and penalizes mismatches in set membership. The likelihood of the entire observation o, $p(o|c)$ can be computed by combining the individual $p(x|c)$ terms (see SI.2.3 for details).

This likelihood function can then be used to update the agent's posterior belief about the contexts it considers viable at decision instance t, given its observation history as,

$$p(c^{(t)}|o^{(1:t)}) = \frac{p(o^{(t)}|c)p(c|o^{(1:t-1)})}{\sum_c^{\mathcal{C}} p(o^{(t)}|c)p(c|o^{(1:t-1)})}, \tag{2.3}$$

To outline a decision theory within this framework, observe that, at decision instant t, a Bayesian agent could represent its prior preference for different world possibilities in the form of a probability distribution over the possible outcomes in $\mathcal{X}$, conditioned on desirability information obtained in earlier decisions, $p(x|c^{(t)}, r^{(1:t-1)})$. New evidence for the desirability of outcomes observed in context $c^{(t)}$ is incorporated using $p(r^{(t)}|x, c^{(t)})$, a distribution encoding the relative desirability information obtained from the environment at the current time step, conditioned on the context in which the information is obtained. This formulation immediately yields the belief update,

$$p(x|c^{(t)}, r^{(t)}) \propto p(r^{(t)}|c^{(t)}, x) \times p(x|c^{(t)}, r^{(1:t-1)}), \tag{2.4}$$

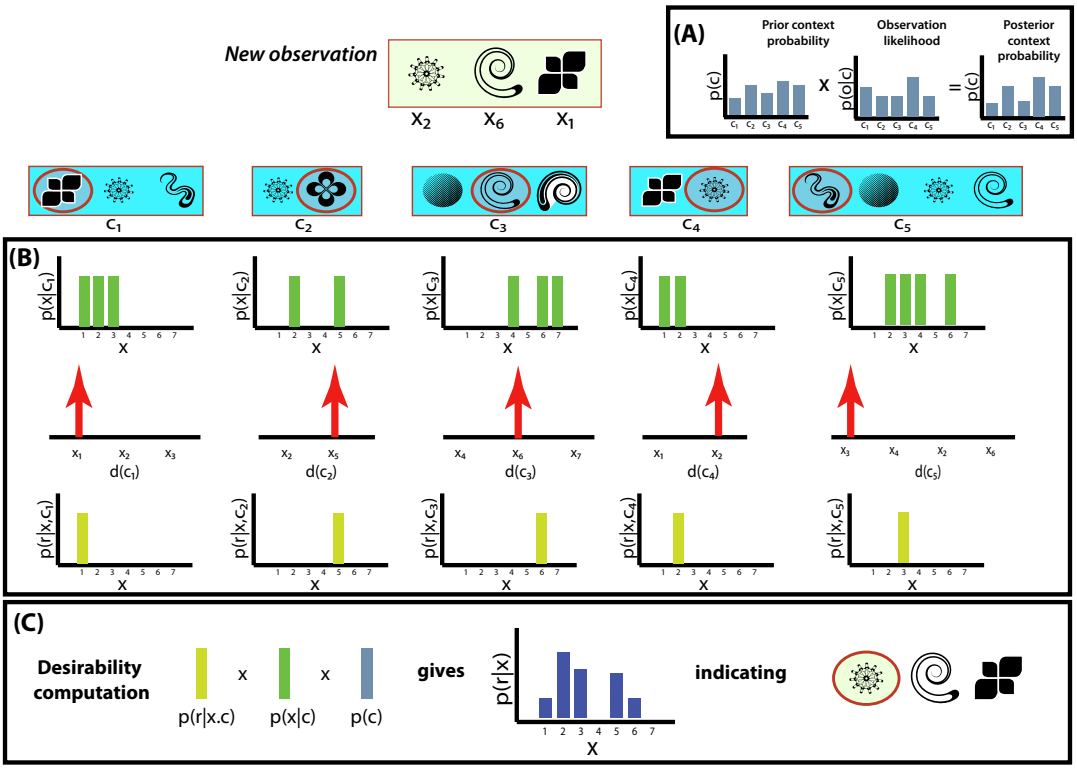


Figure 2.3: A graphical illustration of the relative desirability computation. An observer sees a subset of all possibilities in the world at a given time. The observer **(A)** updates its context probability based on the current observation. Subsequently, **(B)** the observer computes desirability and observation probabilities for the new observation and **(C)** uses these intermediate computations to rationally infer relative desirability values for the new option set.

to obtain a posterior probability encoding the desirability of different possibilities x, while also accounting tractably for the context in which desirability information is obtained at every decision instance. Defining a choice function to select the mode of the posterior belief completes a rational context-sensitive decision theory.

2.2 Results

To demonstrate the value of the relative desirability-based encoding of preferences, in 2.2.1, we describe situations in which the influence of context shifting significantly affects human preference behavior in ways that utility-based decision theories have historically been hard-pressed to explain. Complementarily, in 2.2.2 we characterize conditions under which the relative desirability framework yields predictions of choice behavior equivalent to that predicted by ordinal utility theories, and hence, is an equivalent representation for encoding preferences. In 2.2.3, we demonstrate how a relative desirability-based encoding of preference behaves identically with an expected utility computation in choices made under outcome uncertainty, thereby generalizing our framework for utility-free value representation to encompass risky decisions. Finally, in 2.2.4, we demonstrate the representational limitations of this theory in accounting for the effects of probability perceptions in explaining human preferences, thereby delimiting the scope of its applicability.

2.2.1 Desirability learning explains context effects under choice set variation

In this section, we show how our inductive theory of context-sensitive value inference leads, unsurprisingly, to a simple explanation for the major varieties of context effects seen in behavioral experiments. These are generally enumerated as attraction, similarity, comparison and reference point effects [37]. Interestingly, we find that each of these effects can be described as a special case of the frog legs example, with the specialization arising out of additional assumptions made about the relationship of the new option added to the choice set. Table 2.2, with some abuse of notation, describes this relationship between the effects in set-theoretic terms.

In each of the following cases, an *inductively rational* observer will select the option

Effect name	Description	Assumptions
Frog legs	$c_1 \leftarrow \{X, Y\} \Rightarrow X \succ Y$, $c_2 \leftarrow \{X, Y, Z\} \Rightarrow Y \succ X$	-
Similarity	$c_1 \leftarrow \{X, Y\} \Rightarrow X \succ Y$, $c_2 \leftarrow \{X, Y, Z\} \Rightarrow Y \succ X$	$Z \approx X$
Attraction	$c_1 \leftarrow \{X, Y\} \Rightarrow X \sim Y$, $c_2 \leftarrow \{X, Y, Z\} \Rightarrow X \succ Y$	$X \succ Z$
Reference point	$c_1 \leftarrow \{X, Y\} \Rightarrow X \succ Y$, $c_2 \leftarrow \{X, Y, Z\} \Rightarrow X \succ^{(-)} Y$	$Z \succ X$
Compromise	$c_1 \leftarrow \{X, Y\} \Rightarrow X \succ Y$, $c_2 \leftarrow \{X, Y, Z\} \Rightarrow Y \succ X$	$Y \succ^{(c)} X, Z$

Table 2.2: A unified description of context effects. $\succ$ indicates stochastic preference for one item over another. $\succ^{(c)}$ indicates that the preference in question holds only in some observation contexts. $\succ^{(-)}$ indicates that the preference in question is stochastically weaker than before.

that has a greater relative desirability value, which, in turn, is computed using Equation 2.2.

In the frog legs example, the reversal in preferences is anecdotally explained by the diner originally forming a low opinion of the restaurant's chef, given the paucity of choices on the menu, deciding to pick the safe salmon over a possibly a burnt steak. However, the waiter's presenting frog legs as the daily special suddenly raises the diner's opinion of the chef's abilities, causing him to favor steak. This intuition maps very easily into our framework of choice selection, wherein the diner's partial menu observations o_1 = {steak, salmon} and o_2 = {steak, salmon, frog legs} are associated with two separate contexts c_1 and c_2 of observing the menu $\mathcal{X}$. Bad experiences related to ordering steak in menus typically observed under context c_1 (interpretable as 'cheap restaurants') may be encoded by defining the vector $m = \{1, 0, 0, 0\}$ for c_1 and good experiences ordering steak off menues observed in context c_2 (interpretable as 'upscale restaurants') as $m = \{0, 1, 0, 0\}$ for c_2. Then, by definition, $p(r|\text{salmon}, c_1) > p(r|\text{steak}, c_1)$, while $p(r|\text{salmon}, c_2) < p(r|\text{steak}, c_2)$. For the purposes of this demonstration, let us assume these probability pairs, obtained through the diner's past experiences in restaurants to be $\{0.7, 0.3\}$ and $\{0.3, 0.7\}$ respectively. Now, when the waiter first offers the diner a choice between steak or salmon, the diner computes relative desirabilities using (2.2), where the only context for the observation is {salmon, steak}. Hence, the relative desirabilities of steak and salmon are computed over a single context, and are simply $R(\text{salmon}) = 0.7, R(\text{steak}) = 0.3$. When the diner is next presented with the possibility

of ordering frog legs, he now has two possible contexts to evaluate the desirability of his menu options: {salmon, steak} and {salmon, steak, frog legs}. Based on the sequence of his history of experience with both contexts, the diner will have some posterior belief $p(c) = \{p, 1-p\}$ on the two contexts. Then, the relative desirability of salmon, after having observed frog legs on the menu can be calculated using (2.2) as,

$$\begin{aligned} R(\text{salmon}) &= \frac{p(r|\text{salmon}, c_1)p(\text{salmon}|c_1)p(c_1) + p(r|\text{salmon}, c_2)p(\text{salmon}|c_2)p(c_2)}{p(\text{salmon}|c_1)p(c_1) + p(\text{salmon}|c_2)p(c_2)}, \\ &= \frac{0.7 \times 1 \times p + 0.3 \times 1 \times (1-p)}{1 \times p + 1 \times (1-p)} = 0.7p + 0.3(1-p). \end{aligned}$$

Similarly, we obtain $R(\text{steak}) = 0.3p + 0.7(1-p)$. Clearly, for $1-p > p, R(\text{steak}) > R(\text{salmon})$, and the diner would be rational in switching his preference. Thus, through our inferential machinery, we retrieve the anecdotal explanation for the diner's behavior: if he believes that he is more likely to be in a good restaurant (with probability $(1-p)$) than not, he will prefer steak.

Along identical lines, making reasonable assumptions about the contexts of past observations, our decision framework accomodates parsimonious explanations for each of the other effects detailed in Table 2.2.

In all the cases described below, we assume particular initial levels of relative desirability, implying a particular class of history of desirability observations in the observer's history. It is possible in some cases to find partitions of relative desirability between options that render a particular demonstration invalid. We will point these out where relevant, and interpret them as theoretical substantiations of the observed fragility of these effects to changes in the relative desirabilities of the options in the initial choice set.

Similarity effect

In the similarity effect, given a choice set XY, a subject prefers option X. Now, a third option Z is introduced into the choice set, which is known to be similar to X, and not generally perceived to be clearly superior or inferior to X. In this expanded choice set XYZ, the subject is observed to reverse his preference and select Y.

Because of the similarity between X and Z, evidence for the desirability of X is

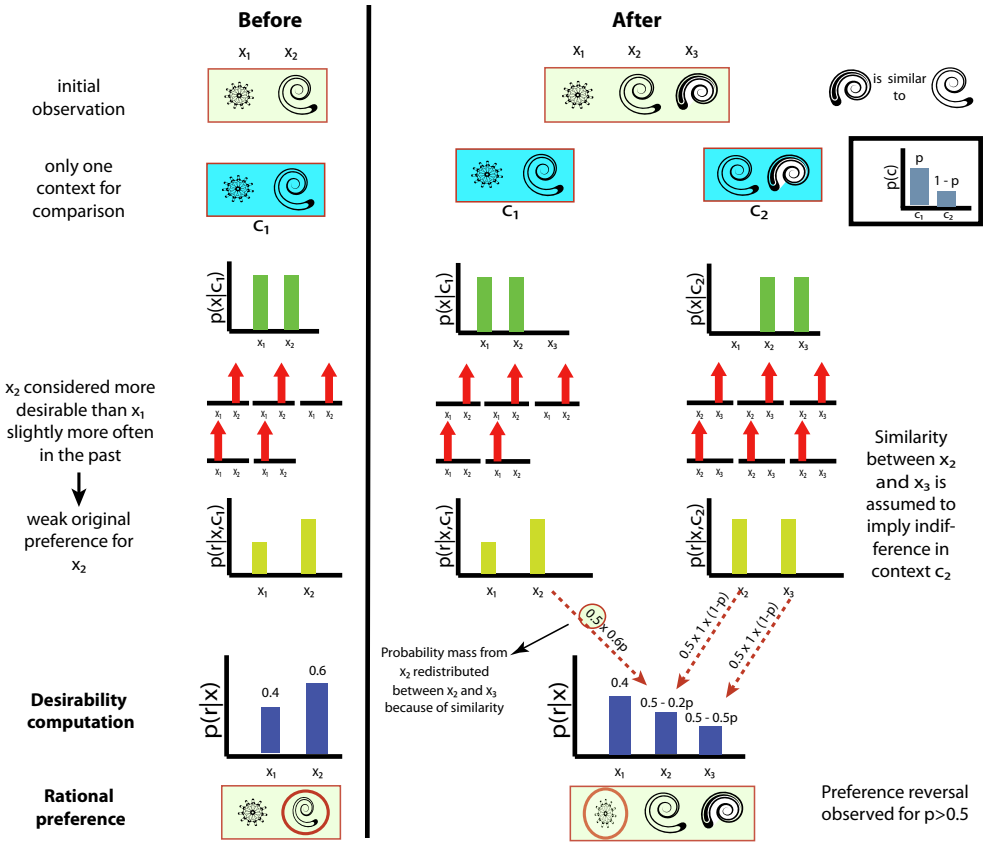


Figure 2.4: Mathematization of the similarity effect in our desirability elicitation framework. Similarity between the existing preferred object and a novel object causes a redistribution of the desirability probability of the original object between them. For a range of desirability and probability values, this leads to a reversal from the original preference.

computed over both observations of X and Z such that,

$$\begin{aligned} R(X) &= \frac{\sum_x^{\{X,Z\}} \sum_1^{10} p(r|x,XY)p(x|XY)p(XY) + \sum_x^{\{X,Z\}} \sum_1^{10} p(r|x,XZ)p(x|XZ)p(XZ)}{\sum_x^{\{X,Z\}} \sum_1^{10} p(x|XY)p(XY) + \sum_x^{\{X,Z\}} \sum_1^{10} p(x|XZ)p(XZ)}, \\ &= \frac{6 \times 1 \times 1 \times p + 0 + 10 \times 1 \times 1 \times (1-p)}{10 \times 1 \times p + 10 \times 1 \times (1-p) + 10 \times 1 \times p + 10 \times 1 \times (1-p)}, \\ &= \mathbf{0.5 - 0.2p}, \end{aligned}$$

while a similar computation for $R(Y)$ yields,

$$\begin{aligned} R(Y) &= \frac{\sum_1^{10} p(r|Y,XY)p(Y|XY)p(XY) + \sum_1^{10} p(r|Y,XZ)p(Y|XZ)p(XZ)}{\sum_1^{10} p(Y|XY)p(XY) + \sum_1^{10} p(Y|XZ)p(XZ)}, \\ &= \frac{4 \times 1 \times 1 \times 0.5 + 0}{10 \times 1 \times 0.5 + 0}, \\ &= \mathbf{0.4}, \end{aligned}$$

and $R(Z) = \mathbf{0.5 - 0.5p}$.

For all values of $p > 1/2$, $R(X) < R(Y)$, resulting in a rational preference reversal $Y \succ X$, suggesting that the similarity effect is sensitive to the extent to which the comparison between the similar objects becomes salient in the choice domain. This dependence has an intuitive explanation. For high values of p, such comparisons will be rare, and cause a preference reversal in the original choice set. However, when the similarity comparison dominates the original choice comparison, the subject generalizes his preference $X \succ Y$ to the new object Z.

Re-calculation using different values for the XY preference (e.g. 9/1 instead of 6/4) also suggests that the similarity effect will disappear in cases where X is clearly preferable to Y, an easily testable prediction from our theory. Further, the inference mechanism allows us to also predict that the similarity effect will return in such cases with the introduction of yet more items Z' similar to X into the choice set, in line with empirical observations on the similarity effect.

Attraction effect

In the attraction effect, given a set of choices, {X,Y}, the subject is originally seen to be indifferent between the two options. However, when a third option Z that is similar to,

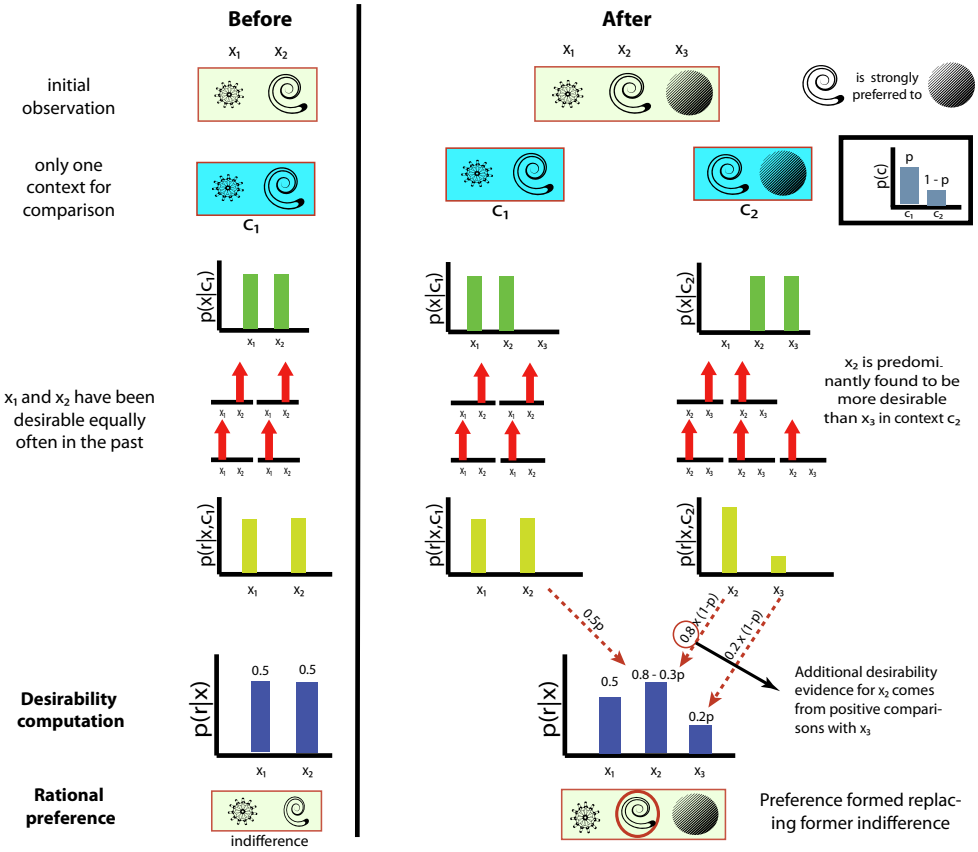


Figure 2.5: Mathematization of the attraction effect in our desirability elicitation framework. Positive comparisons of an object wthin the existing choice set with a novel object increases its relative desirability, thereby inducing a preference for it in place of indifference in the original choice set.

but clearly inferior to option X is introduced into the choice set, the subject's preference switches to prefer option Y.

In the extended choice set regime, the desirability computation for X yields,

$$\begin{aligned} R(X) &= \frac{\sum_1^{10} p(r|X,XY)p(X|XY)p(XY) + \sum_1^{10} p(r|X,XZ)p(X|XZ)p(XZ)}{\sum_1^{10} p(X|XY)p(XY) + \sum_1^{10} p(X|XZ)p(XZ)}, \\ &= \frac{5\times 1\times 1\times p + 8\times 1\times 1\times (1-p)}{10\times 1\times p + 10\times 1\times (1-p)}, \\ &= \mathbf{0.8 - 0.3p}, \end{aligned}$$

while a similar computation for Y yields,

$$\begin{aligned} R(Y) &= \frac{\sum_1^{10} p(r|Y,XY)p(Y|XY)p(XY) + \sum_1^{10} p(r|Y,XZ)p(Y|XZ)p(XZ)}{\sum_1^{10} p(Y|XY)p(XY) + \sum_1^{10} p(Y|XZ)p(XZ)}, \\ &= \frac{5\times 1\times 1\times 0.5 + 0}{10\times 1\times 0.5 + 0}, \\ &= \mathbf{0.5}. \end{aligned}$$

p being a probability with non-zero support for both contexts $p \in (0,1) \Rightarrow R(X) > R(Y)$, resulting in the establishment of a rational preference $X \succ Y$ in place of the earlier indifference. This conclusion is expected to hold for any possible combinations of XZ preferences that clearly favor X. Furthermore, our inferential process in this setup also predicts that an item that is originally equivalent in desirability to another item that is then found to be inferior to a third item in a separate choice set will be subsequently found preferred in the original comparison. Interestingly, this secondary prediction has been verified in human infants and capuchin monkeys by [38]. While Egan et al consider their findings to be evidence for cognitive dissonance, the results from their experimental task are clearly interpretable as an (un)attraction effect, and hence, compatible with our explanation.

Reference point effect

The reference point effect has been used to explain many divergent sets of phenomena in the behavioral economics literature. For our demonstration, we restrict ourselves to explaining the results of a particular experiment on human subjects due to Vlaev et al [39], where subjects paid money to avoid forthcoming electric shocks of three different

intensities, low, medium and high. The researchers found that subjects consistently paid more money to avoid pains that were greater than others in their recent history. In two sets of experiments, one where low shocks were mixed with medium shocks and one where medium shocks were mixed with higher ones, it was found that subjects paid much more money to buy out of medium shocks in the first condition than the second. Essentially, their evaluation of the undesirability of a particular magnitude of pain was contingent on the set of pain options that they were forced to choose between. In the context-presentation framework, this can be posed as a problem where the subject is first offered the choice set LM, followed by further exposure to the choice set MH. Assuming that the subject always prefers the option that provides the lesser amount of pain, their evaluation of relative (un)desirability of the medium option after experience with the choice set LM will be $R(M) = 1$. Upon further experience with MH, the new desirability of M can be computed as,

$$\begin{aligned} R(M) &= \frac{\sum_1^{10} p(r|M, LM)p(M|LM)p(LM) + \sum_1^{10} p(r|M, MH)p(M|MH)p(MH)}{\sum_1^{10} p(M|LM)p(LM) + \sum_1^{10} p(M|MH)p(MH)}, \\ &= \frac{10p + 0}{10}, \\ &= \mathbf{p}, \end{aligned}$$

which, being less than 1, implies that the (un)desirability of M reduces after exposure to a higher degree of pain. This observation, while almost trite on surface, has eluded the descriptive abilities of utility function approaches of measuring value, as described comprehensively in [28]. This effect is expected to remain stable for any choice of relative desirability frequency that respects the intuition that relatively lower levels of pain are more preferable.

Compromise effect

In the compromise effect, given a set of choices, {X,Y}, a subject prefers option X. Introduction of a third option Z leads to the development of two different ways of evaluating the desirability of any of the three items, resulting in situations where X may be strongly preferred to Z along one axis of measurement and strongly dominated by Z along the other. In standard descriptions of this effect, these different ways of evaluation are regarded as attributes, leading to a simple description of the problem in

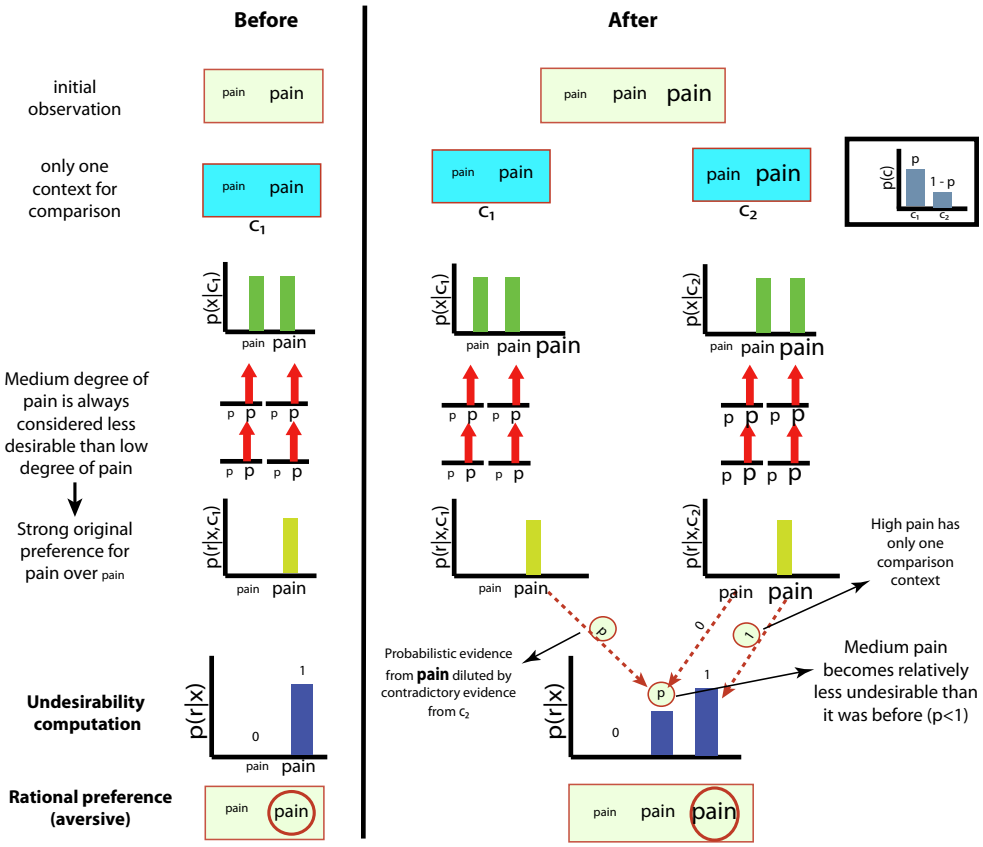


Figure 2.6: Mathematization of the reference point effect in our desirability elicitation framework. Contradictory evidence from two sets of comparisons reduces the (un)desirability of the intermediate option, resulting in the agent evincing relative evaluation. Note that we use aversive desirability pointers pointing to the *least* desirable options in this setting. In general, both aspirational and aversive reactions could arise for the same set of stimuli. By formulating two separate accessibility relations for these two modes of engagement, our inductive framework allows for the possibility of the same object being both extremely desirable and extremely undesirable simultaneously to an observer.

the framework of multi-attribute utility theory. For our current purpose, we achieve the same purpose notationally by considering XY and YX to be two different observation contexts representing possibilities that always co-occur, but are not always evaluated identically. Now, as in the earlier examples, at the time of the first observation, the only possible context is XY; an observation history containing 7 preferences for X and 3 for Y results in a relative desirability calculation, $R(X) = 0.7, R(Y) = 0.3$.

Introduction of the third option, however, results in the (recalled) feasibility of six different contexts, which we index in $\mathcal{C} = \{XY, YX, YZ, ZY, ZX, XZ\}$. By the premise of the compromise effect setup, in the history of observing XZ, X is preferred 8 times, while Z is preferred twice, while in observing ZX, these numbers are reversed. Say observing 10 instances of YX yields 8 preferences for Y and 2 for X. 10 instances of YZ yield (6Y,4Z) while ZY yields (6Z,4Y). Since the contexts ij and ji are indistinguishable as observable choice sets, they occur with the same sample frequency. Thus, we can assume a posterior belief on six contexts, $\{p_1, p_1, p_2, p_2, p_3, p_3\}$. Then, upon observing XYZ, the desirability computation for X yields,

$$\begin{aligned} R(X) &= \frac{\sum_c^{\mathcal{C}} \sum_1^{10} p(r|X,c)p(X|c)p(c)}{\sum_c^{\mathcal{C}} \sum_1^{10} p(X|c)p(c)}, \\ &= \frac{7p_1 + 2p_1 + 0 + 0 + 2p_3 + 8p_3}{10p_1 + 10p_1 + 0 + 0 + 10p_3 + 10p_3}, \\ &= 0.05\frac{9p_1 + 10p_3}{p_1 + p_3}. \end{aligned}$$

A similar computation for $R(Y)$ yields,

$$\begin{aligned} R(Y) &= \frac{\sum_c^{\mathcal{C}} \sum_1^{10} p(r|Y,c)p(Y|c)p(c)}{\sum_c^{\mathcal{C}} \sum_1^{10} p(Y|c)p(c)}, \\ &= \frac{3p_1 + 8p_1 + 2p_2 + 8p_2 + 0 + 0}{10p_1 + 10p_1 + 10p_2 + 10p_2 + 0 + 0}, \\ &= 0.05\frac{11p_1 + 10p_2}{p_1 + p_2}. \end{aligned}$$

Setting $p_2 = p_3$ is equivalent to assuming Y and X both have equal histories of comparisons with the new option, which, while never a stated condition for observing the compromise effect, is not *prima facie* unreasonable. Doing so immediately forces $R(X) < R(Y)$, rendering the preference $Y \succ X$ rational. The compromise effect has many more assumptions about relative preference frequencies than the attraction and

similarity effect descriptions, rendering a comprehensive analysis intractable. It is clear, however, that assuming a symmetric relationship between the XY and YX preferences, as we do in all the other cases, partially breaks the compromise effect, by rendering $X \sim Y$. Hence, we predict that a necessary requirement for the compromise effect to hold is for option Y to be more clearly preferable than option X along the new axis of evaluation introduced by inclusion of Z in the choice set.

Interpretation of results

We emphasize the simplicity with which an inductive explanation of each of the studied effects arises in our representational framework. The attraction effect arises through the introduction of additional evidence of the desirability of one of the options from a new context, causing the relative desirability of this particular option to rise. The similarity effect is elicited simply as a property of division of probability among multiple similar options, resulting in reduced desirabiliy of the previously superior option. The compromise effect arises through a combination of reduction in the desirability of the superior option through negative comparions with the new item and increase in the desirability of the formerly inferior item through positive comparisons with the new item, and that this inference occurs automatically in our framework assuming equal history of comparisons between the existing choice set items and the new item.

Reference point effects have typically not been associated with explicit studies of context variation, and may in fact be used to reference a number of behavior patterns that do not satisfy the definition we provide in Table 2.2. Our definition of the reference point effect is particularized to explain data on pain perception collected by [39], demonstrating relativity in evaluation of objectively identical pain conditions depending on the magnitude of alternatively experienced pain conditions. In concord with empirical observation, we show that the relative (un)desirability of an intermediate pain option reduces upon the experience of greater pain, a simple demonstration of prospect relativity that utility-based accounts of value cannot match.

Competing hypotheses that seek to explain these behaviors are either descriptive and static (e.g. quantum cognition [40]), normative and static, (e.g. extended discrete choice models ([41] provides a recent review), componential context theory [25]) or descriptive and dynamic, (specifically, decision field theory [24]). In contrast, our approach not only

takes a dynamic inductive view of value elicitation, it retains a normativity criterion (Bayes rationality) for falsifying observed predictions, a standard that is expected of any *rational* model of decision-making [42].

2.2.2 Desirability learning generalizes utility function representations of preferences

It could be conjectured that the relative desirability indicator d will be an inadequate representation of preference information compared with scalar utility signals assigned to each world possibility, which would leave open the possibility that we may have retrieved a context-sensitive decision theory at the expense of theoretical assurance of rational choice selection, as has been the case in many previous attempts cited above. Were this conjecture to be true, it would severely limit the scope and applicability of our proposal. To anticipate this objection, we theoretically prove that our framework reduces to the standard utility-based representation of preferences under equivalent epistemic conditions, showing that our theory retains equivalent rational representational ability as utility theory in simple, and simply extends this representational ability to explain preference behaviors that utility theory can't.

What does it mean for a measure to represent preference information? To show that a utility function u completely represents a preference relation on $\mathcal{X}$ it is sufficient [19] to show that, $\forall x_1, x_2 \in \mathcal{X}, x_1 \succ x_2 \Leftrightarrow u(x_1) > u(x_2)$. Hence, equivalently, to show that our measure of relative desirability R also completely represents preference information, it should be sufficient to show that, for any two possibilities $x_i, x_j \in \mathcal{X}$, and for any observation context c

$$x_i \succ x_j \Leftrightarrow R(x_i) > R(x_j). \tag{2.5}$$

Naturally, this will not be true in general for context-sensitive agents, since our framework specifically allows for preference reversals across multiple contexts, immediately rendering the LHS condition $x \succ y$ insufficiently descriptive of preference relations in general. We find (see SI.2.3 for the proof) that (2.5) holds at decision instant t under three conditions[3] , enumerated in Box 2.

[3] In condition **(III)**, the notation $\mathcal{C}_{i\setminus j}$ references the subset of all observed contexts that contain x_i but not x_j.

Conditions for representation equivalence between relative desirabilities and utilities

(I) **Context consistency:** $\exists c \in \mathcal{C}, s.t.\ x_i \succ x_j \Rightarrow x_i \succ x_j \forall c \in \mathcal{C}_{ij}, \{x_i, x_j\} \in \mathcal{C}_{ij} \subseteq \mathcal{C}$.

(II) **Transitivity between contexts:** if $x_i \succ x_j$ in c_1 and $x_j \succ x_k$ in $c_2, \forall c \in \mathcal{C}, x_i \succ x_k$.

(III) **Symmetry in context observability:** $\forall x_i, x_j \in \mathcal{X}, \lim_{t\to\infty} |\mathcal{C}^{(t)}_{i\setminus j}| = |\mathcal{C}^{(t)}_{j\setminus i}|$.

Of these three assumptions, **(I)** and **(II)** simply define a stable preference relation across observation contexts and find exact counterparts in the completeness and transitivity assumptions necessary for representing preferences using ordinal utility functions. **(III)**, the only additional assumption we require, ensures that the agent's history of partial observations of the environment does not contain any useful information. The restriction of infinite data observability, while stringent and putatively implausible, actually uncovers an underlying epistemological assumption of utility theory, viz. that utility/desirability values can somehow be obtained *directly* from the environment. Any inference based preference elicitation procedure will therefore necessarily need infinite data to attain formal equivalence with the utility representation. Finally, we point out that our equivalence result does not require us to assume continuity or the equivalent Archimedean property to encode preferences, as required in ordinal utility definitions. This is because the continuity assumption is required as a technical condition in mapping a discrete mathematical object (a preference relation) to a continuous utility function. Since relative desirability is defined constructively on $Q \subseteq \mathbb{Q}, |Q| < \infty$, a continuity assumption is not needed.

2.2.3 Desirability learning generalizes expected utility representations of value in risky decisions

In this section, we describe how our inductively rational value inference can be used to evaluate a sense of desirability that is equivalent to the expected utility paradigm generally used for modeling choices under uncertainty. While the earlier examples we discuss have focused on the advantages of using context-sensitive value inference in

problems where outcome observability is partial because of *deterministic* aspects of the environment, this framework can be easily extended to encompass problems where outcome observability is partial because of *stochastic* aspects of the environment, viz. choice problems involving risk.

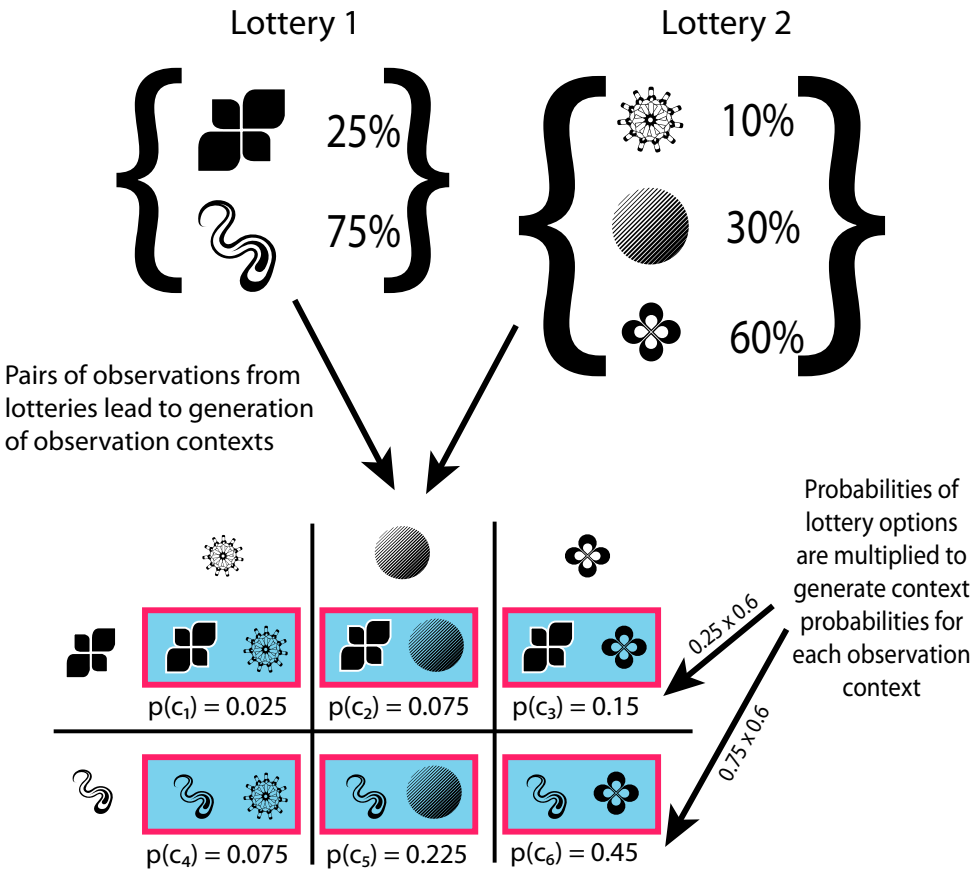


Figure 2.7: While we have no explicit representation of uncertainty in our framework, we observe that partial observability of options generalizes uncertainty, giving us preferences that encode the expected utilities of risky gambles.

Consider an observer who has a choice between two lotteries M and N (see Figure 2.7 for a schematic description), each of which can be indexed by a set of tuples representing the payoffs and payoff probabilities respectively for each option of each lottery, .e.g.

$M = \{(m_1, p_1), (m_2, p_2), (m_3, p_3)\}$ and $N = \{(n_1, q_1), (n_2, q_2), (n_3, q_3)\}$, where $\sum p_i = \sum q_i = 1$. Such risky lotteries map on to our framework with the assumption that binary comparisons between each of the options in these lotteries will yield observations that predict distinct comparative contexts. Thus, for the notional choice of M and N above, we obtain $3 \times 2 = 6$ comparative contexts spanning all the possibilities $\{m_i \times n_j\}$. For comparisons between equal quantities, we assume $p(r|x, c) = 0.5$. For all other comparisons, $p(r|x, c) = 1$ for the option indicating the larger value. Finally, assuming that observations from each lottery are sampled independently, the probability of a particular binary comparison can be computed as $p(\{m_i, n_j\}) = p_i \times q_j$, with the probability values obtained from the lottery definitions.

For instance, consider a choice between two options A and B, one yielding \$20 10% of the time and \$0 the rest of the time, and the other yielding \$50 5% of the time and \$0 the rest of the time. Such standard choice tasks map on to our framework with the assumption that binary comparisons between the three quantities {0, 20, 50} yield observations that predict past comparative contexts. Thus, in assessing the relative desirability of these two options, the set of comparative contexts $\mathbb{C} = \{\{0, 0\}, \{20, 0\}, \{0, 50\}, \{20, 50\}\}$ are activated in the subject's inductive recollection. Then, the relative desirabilities of these options can be computed as,

$$\begin{aligned} R(A) &= \frac{\sum_c^{\mathbb{C}} p(r|A, c)p(A|c)p(c)}{\sum_c^{\mathbb{C}} p(A|c)p(c)} \\ &= \frac{0.5 \times 1 \times 0.9 \times 0.95 + 1 \times 1 \times 0.1 \times 0.95 + 0 \times 1 \times 0.9 \times 0.05 + 0 \times 1 \times 0.1 \times 0.05}{1 \times 0.9 \times 0.95 + 1 \times 0.1 \times 0.95 + 1 \times 0.9 \times 0.05 + 1 \times 0.1 \times 0.05}, \\ &= \mathbf{0.5225}, \end{aligned}$$

and,

$$\begin{aligned} R(B) &= \frac{\sum_c^{\mathbb{C}} p(r|B, c)p(B|c)p(c)}{\sum_c^{\mathbb{C}} p(B|c)p(c)} \\ &= \frac{0.5 \times 1 \times 0.9 \times 0.95 + 0 \times 1 \times 0.1 \times 0.95 + 1 \times 1 \times 0.9 \times 0.05 + 1 \times 1 \times 0.1 \times 0.05}{1 \times 0.9 \times 0.95 + 1 \times 0.1 \times 0.95 + 1 \times 0.9 \times 0.05 + 1 \times 0.1 \times 0.05}, \\ &= \mathbf{0.4775}, \end{aligned}$$

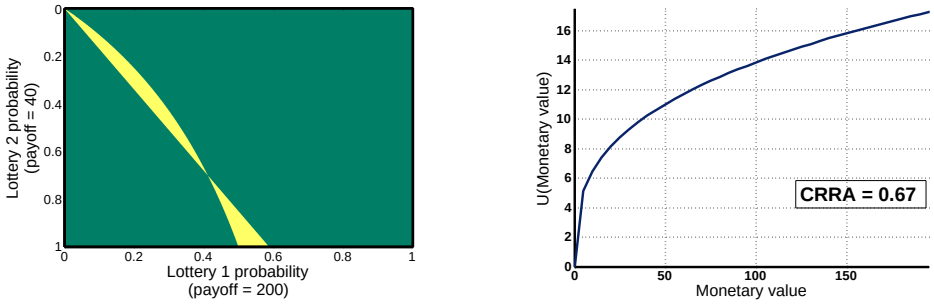
resulting in a preference for option A over option B. While such a preference would not be rational under an expected value decision theory, key aspects of the decision

problem are left unspecified in our present framework, e.g. the utility function that maps cardinal monetary values to hedonic desirabilities and/or the subjective weights assigned to different probability values via probability weighting functions. Here, we simply demonstrate the ability of our inductive value inference scheme to *represent* expected utilities obtained in decisions under uncertainty.

As [43, 3] demonstrate in their 'decision by sampling' research, comparative measures of value yield relative preferences for monetary attribute values that are indistinguishable from standard psycho-economic formulations of these quantities given a realistic prior distribution of attribute values. In Figure 2.8, we report results from a simulation study of relative pricing of two risky options, where we find that our inductive value inference framework also consistently results in risk averse preferences, hence indicating diminishing marginal utilities. Our simulated task is identical with the multiple price list (MPL) method proposed by [44] in the context of value elicitation, which has subsequently become a standard assessment for hypothetical pricing [45]. In standard MPL, subjects are asked to select between two lotteries, one paying a higher amount with a lower probability, and the other paying a lower amount with a higher probability. For sufficiently low and high probabilities respectively, all subjects should prefer the second lottery to the first. In the task, the probabilities of both lotteries are increased and decreased respectively until subjects switch their preference to the first lottery. The point of switching is considered to be where the implied expected utility of the two lotteries appears equal to subjects. Assuming that probabilities are observed and processed perfectly, this implies that $p_1U_1 = p_2U_2$, at $\{p_1, p_2\}$, an indifference point in probabilities resulting in an estimate of the ratio of utilities for different monetary options.

The MPL method is known [46] to have three disadvantages: (i) it can only elicit valuations within intervals specified by the size of the magnitude of probability change between iterations, (ii) it may elide the impact of inconsistent switching behavior of subjects and (iii) it elides the possible impact of psychological anchoring to the middle of the sequence in subjects. Our simulation setting allows us to sidestep each of these problems. First, we can make intervals arbitrarily small in our simulation, so that the standard interval valuation response of MPL converges to a point response. Second, we test switching not just for one sequence of increasing and decreasing probabilities,

but across the entire spectrum of possible probability comparisons, viz. gambles corresponding to all possible probability combinations for both lotteries are compared with each other. The third problem is not salient to computational agents.



(a) Comparison of predictions of relative desirability-based valuation in a multiple price list task to those from expected utility theory with a fixed CRRA. The numerical value of the CRRA is selected to minimize the disagreement in predictions between the two models.

(b) Implied isoelastic utility function from simulated multiple price list task. The best fit numerical value of CRRA increases with the ratio of the lottery payoffs, but remains within the range (0.2,0.9) for payoff ratios varying between 2 to 200.

Figure 2.8: Relative desirability valuation leads to pricing risky gambles in a manner behaviorally identical to rational pricing by agents with concave utility functions. The numerical value of the best CRRA fit lies well within the range of CRRA values observed in human data [45].

Performing these comparisons generates a prediction about the rational preference one would expect an agent following our inference method to hold for each point on the $p_1 \times p_2$ grid. Comparing the predictions of such an agent with those of an expected utility maximizing agent yields regions in this space where the predictions of both theories are in congruence (shown in green in Figure 2.8(a)), and regions where they conflict (shown in yellow in Figure 2.8(a)). Since both agents have complete access to

probability information. We find the implicit relative risk aversion of the inferential agent by assuming that its utility function is isoelastic, i.e.

$$U(\text{lottery value}) = \frac{\text{lottery value}^{(1-\rho)}}{1-\rho},$$

where ρ is the CRRA. Using this equation, we supply both inferential and expected utility-based simulated agents in the MPL setup with implied utilities corresponding to the lottery payoffs for multiple values of ρ. The value that minimizes the discrepancy between the inferential and expected utility models' predictions is an estimate of the implied CRRA from the data, optimal in an L_1 regression sense. Figure 2.8(b) plots the utility function that best fits the data obtained from the simulated MPL task. The CRRA estimate remains positive for any choice of lottery payoffs such that the ratio of the greater to the smaller payoff is greater than 1.5[4] , verifying the natural emergence of diminishing marginal utility from our theory. Unlike the demonstration of marginal utility from relative desirability pointers in [3, 28], our method does not require the inducement of an informative prior, specifying that larger numbers are encountered less often, on the space of monetary values. In our case, the underestimation of the larger option emerges simply from the absence of ratio-based comparisons, and therefore, provides a more parsimonious explanation for the predominant presence of relative risk aversion in human subjects [45].

2.2.4 Desirability learning cannot explain probability distortion effects in risky decisions

While our definition of inductive rationality departs from standard economic assumptions in useful ways, capturing a broad class of context variation effects, it is clear that our generalization still does not entirely capture the entire canvas of human economic behavior. This is because our approach is dependent on the existence of *rational* and empirically accessible probabilities, which is behaviorally unjustifiable. Human probability assessment presents analogous difficulties to value inference, resisting axiomatization through a stream of probability paradoxes and biases documented in the literature (see e.g. [47] for a recent review). Our theory cannot account for distortions in probability

[4] The estimate drops below zero for ratios below this range due to artifacts in the fitting procedure.

judgment, and hence, is incapable of explaining economic behavior that emerges through them. As a concrete example, we demonstrate the inability of inductive rationality, as presently defined, to explain the famous Allais paradox (described in Table 2.3).

1A	**1B**	**2A**	**2B**
\$1000, 100% chance	\$1000, 89% chance \$0, 1% chance \$5000, 10% chance	\$0, 89% chance \$1000, 11% chance	\$0, 90% chance \$5000, 10% chance

Table 2.3: The Allais paradox. Subjects that prefer option 1A to option 1B must rationally prefer option 2A to option 2B. However, empirical data shows that human subjects tend to prefer option 1A to 1B and option 2B to 2A in violation of normative expected utility maximization.

The Allais paradox maps on to our theory through the assumption that binary comparisons between the three quantities {0, 1000, 5000} yield observations that predict past comparative contexts. For instance, in contemplating the choice between options 2A and 2B, the comparative contexts {0, 0}, {1, 0} and {1, 5} are activated in the subject's inductive recollection, while the choice between 1A and 1B activates contexts {1, 1}, {1, 0} and {1, 5}. Along the lines of our earlier demonstrations, we can compute relative desirabilities as follows,

$$\begin{aligned} R(2A) &= \frac{p(r|2A,\{0,0\})p(2A|\{0,0\})p(\{0,0\}) + p(r|2A,\{1,0\})p(2A|\{1,0\})p(\{1,0\}) + p(r|2A,\{0,5\})p(2A|\{}{p(2A|\{0,0\})p(\{0,0\}) + p(2A|\{1,0\})p(\{1,0\}) + p(2A|\{0,5\})p(\{0,5\})} \\ &= \frac{0.5\times 1\times 0.89\times 0.90 + 1\times 1\times 0.11\times 0.90 + 0\times 1\times 0.89\times 0.1}{1\times 0.89\times 0.90 + 1\times 0.11\times 0.90 + 1\times 0.89\times 0.1}, \\ &= \mathbf{0.4995}, \end{aligned}$$

and,

$$\begin{aligned} R(2B) &= \frac{p(r|2B,\{0,0\})p(2B|\{0,0\})p(\{0,0\}) + p(r|2B,\{1,0\})p(2B|\{1,0\})p(\{1,0\}) + p(r|2B,\{0,5\})p(2B|}{p(2B|\{0,0\})p(\{0,0\}) + p(2B|\{1,0\})p(\{1,0\}) + p(2B|\{0,5\})p(\{0,5\})} \\ &= \frac{0.5\times 1\times 0.89\times 0.90 + 1\times 1\times 0.11\times 0.90 + 0\times 1\times 0.89\times 0.1}{1\times 0.89\times 0.90 + 1\times 0.11\times 0.90 + 1\times 0.89\times 0.1}, \\ &= \mathbf{0.5005}, \end{aligned}$$

which indicates that preferring option 2B is rational.

Performing a similar computation for the first gamble yields,

$$\begin{aligned} R(1A) &= 0.5 \times 1 \times 0.89 + 1 \times 1 \times \overbrace{\mathbf{0.01}} + 0 \times 1 \times 0.1, \\ &= \mathbf{0.455}, \end{aligned}$$

and,

$$\begin{aligned} R(1B) &= 0.5 \times 1 \times 0.89 + 0 \times 1 \times \overbrace{\mathbf{0.01}} + 1 \times 1 \times 0.1, \\ &= \mathbf{0.545}, \end{aligned}$$

yielding the conclusion that preferring option 1B is preferable, along classical utility maximization expectations, but *contra* the observed behavior of Allais paradox subjects. It is instructive to here recall the classic behavioral studies of prospect theory [48] which have consistently shown that human probability estimates systematically deviate from empirically normative standards, specifically by overestimating extremely low and under-estimating extremely high probabilities [32]. An over-estimate of greater than 10 for the 1% chance of obtaining nothing (highlighted above) in gamble 1B would harmonize the relative desirability calculation with the Allais paradox's predictions[5] . However, such an *ad hoc* addition to our framework is unsatisfactory, and suggests that a more rigorous investigation of subjective probability is necessary to supplement our inductive value inference scheme. In the interim, however, understanding the limitations that probability inference places upon our theory allows us to demarcate the class of behaviors for which it will yield reasonable predictions. In general, this will be precisely the class of behaviors where the **frequency** of change in outcome or context observation is not a salient aspect of the decision-making process. This distinction also clarifies the contribution of our paper: **we have developed an inductive replacement for *value*, not decisions**.

2.3 Discussion

In this work, we have developed a theory of inductively rational elicitation of relative preferences of subjects, based on their history of choice availability and selection. Our

[5] Such an apparently drastic over-estimate of 1% empirical risk is empirically documented in human subject behavior [47]

results characterize conditions of the environment wherein it is appropriate for modelers to describe relative preferences as scalar utilities, and illustrate the importance of option availability in supplying auxiliary information about options under less restrictive environment conditions, leading to simple explanations of the major categories of context effects described in the literature. Ultimately, inferred relative desirability is proposed as a theoretically sound replacement for static utility assignments in choice modeling.

While existing preference elicitation techniques [49, 50, 51] use traditional axiomatic definitions of rationality to restrict the space of preferences they search within, our method infers relative preferences in the most general sense, with rationality imposed via the process of value inference itself. Thus, as we show in 2.2.1, our model yields predictions in line with human behavior exhibiting preferences that standard utility axioms would characterize as intransitive and/or incomplete. Such behavior is irrational from the standpoint of economic rationality, but is subsumed within the inductive sense of rationality implied by our value inference methodology.

Throughout this exegesis, we have encountered three different representations of choice preferences: relative (ordinal) utilities, absolute (cardinal) utilities and our own proposal, viz. relative desirability. Each representation leads to a slightly different definition of rationality, so that, assuming a rational set selection function σ in each case we have,

- **Economic rationality:** $x \in \sigma(\mathcal{X}) \Rightarrow \nexists y \in \mathcal{X}, s.t.\ y \succ x$, predominantly used in human preference modeling in neoclassical economics [19]], e.g. discrete choice modeling [52].
- **VNM-rationality:** $x \in \sigma(\mathcal{X}) \Rightarrow \nexists y \in \mathcal{X}, s.t.\ u(y) > u(x)$, predominantly used in studying decision-making under risk [53], e.g. reinforcement learning [8].
- **Inductive rationality:** $x \in \sigma(\mathcal{X}) \Rightarrow \nexists y \in \mathcal{X},\ s.t.\ R(y, \{H\}) > R(x, \{H\})$, which we have proposed. The term $\{H\}$ here is shorthand for $\{o_1, o_2, \cdots, o_{t-1}\}, \{r_1, r_2, \cdots r_{t-1}\}$, the entire history of choice set and relative desirability observations made by an agent leading up to the current decision instance.

Inductive rationality simply claims that value inference with the same history of

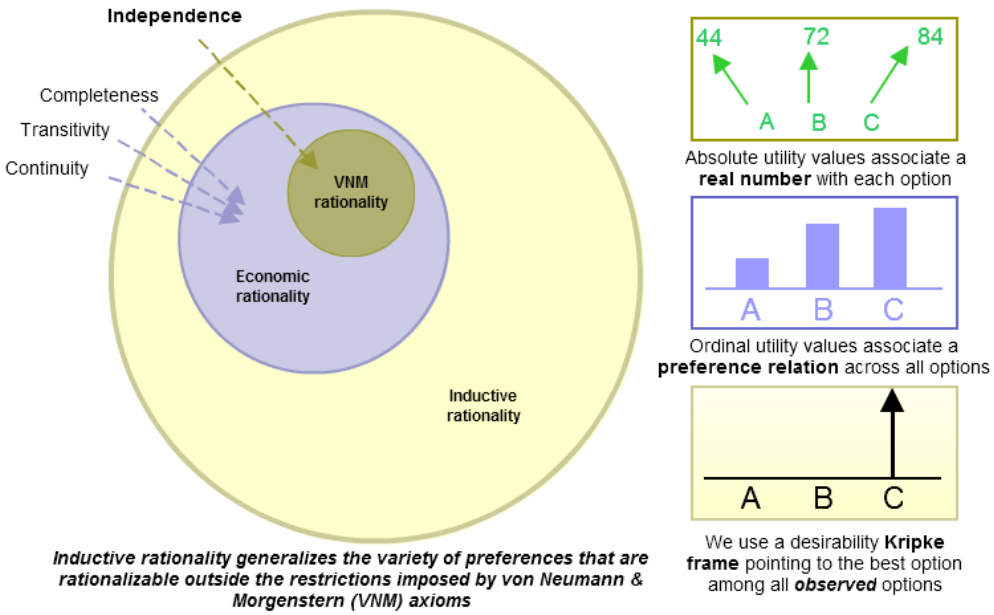


Figure 2.9: A schematic overview of the contribution of this paper. Inductive, or Bayes rationality generalizes existing definitions of rationality while requiring parsimonious epistemic assumptions about human value judgment.

partial observations will lead to a consistent preference for a particular option in discrete choice settings. In Section 2.2.2, we have shown conditions on choice set observations under which inductive rationality will be equivalent to economic rationality. VNM-rationality is a further specialization of economic rationality, valid for preference relations that, in addition to being complete, transitive and continuous (as required for economic preferences representable via ordinal utilities) also satisfy an independence of irrelevant attributes (IIA) assumption [34]. Inductive rationality specializes to economic rationality once we instantiate the underlying intuitions behing the completeness and transitivity assumptions in a context-sensitive preference inference theory. Therefore, rational value inference in the form we propose can formally replace static assumptions about preference orderings in microeconomic models that currently exclusively use ordinal utilities [19]. As such, context-sensitive preference elicitation is immediately useful for the nascent agent-based economic modeling paradigm as well as in dynamic stochastic general equilibrium models of economic behavior. Further work is necessary to develop a context-sensitive equivalent of the IIA assumption, which is necessary for our system to be directly useful in modeling decision-making behaviors under uncertainty. However, even in its current form, our inference model can be used in conjunction with existing 'inverse planning' models of utility elicitation from choice data [54] that infer absolute utilities from choice data using extraneous constraints on the form of the utility function from the environment. In such a synthesis, our model could generate a preference relation sensitive to action set observability, which inverse planning models could use along with additional information from the environment to generate absolute utilities that account for observational biases in the agent's history.

A philosophically astute reader will point out a subtle flaw in our inferential definition of rationality. Namely, while we assume an intuitive notion of partial observability of the world, in practice, our agents compile desirability statistics on the set of all possibilities, irrespective of whether they have ever been observed, a problem that is rooted in an inherent limitation of Bayesian epistemology of being restricted to computing probabilities over a fixed set of hypotheses. How can a desirability representation that assumes that observers maintain probabilistic preferences over all possible states of the world be more epistemologically realistic than one that assumes that observers

maintain scalar utility values over the same state space[6] ? As a partial response to this criticism, we point out that we do not require an ontic commitment to the computation of joint probability distributions on all $x \in \mathcal{X}$. In practice, it is likely that Bayesian computations are implemented in the brain via sampling schemes that, in hierarchical formulations, allow approximating information of the joint distribution as a set of the most likely marginals (in our case, relative desirability in *typical* observation contexts). Neural implementations of such sampling schemes have been proposed in the recent cognitive science literature [55]. Devising a sampling scheme that matches the intuition of context retrieval from memory to supplement our value-inference scheme presents a promising direction for future research.

Another straightforward extension of our framework would imbue observable world possibilities with *attributes*, resulting in the possibility of deriving a more general definition of contexts as clusters in the space of attributes. Such an extension would result in the possibility of transferring preferences to entirely new possibilities, allowing the set $\mathcal{X}$ to be modified dynamically, which would further address the epistemological criticism above. Even further, such an extension maps directly to the intuition of value inference resulting from organisms' monitoring of internal need states, here modeled as attributes. Canini's recent modeling of transfer learning using hierarchical Dirichlet processes [56] provides most of the mathematical apparatus required to perform such an extension, making this a promising direction for future work in our project.

In conclusion, it has long been recognized that state-specific utility representations of the desirability of options are insufficient to capture the rich variety of systematic behavior patterns that humans exhibit. In this paper, we show that reformulating the atomic unit of desirability as a context-sensitive 'pointer' to the best option in the observed set recovers a rational way of representing desirability in a manner sufficiently powerful to describe a broad range of context effects in decisions. Since it is likely that preferences for options do not exist *a priori* and are induced via experience, our present proposal is expected to approximate the true mechanisms for the emergence of context-sensitive preference variation better than alternative static theories, while retaining normativity criteria missing in alternative dynamic accounts. Replacing option-specific

[6] One could argue that we are essentially observing the state space (to be able to index using its membership), but pretending to not observe it.

utilities with relative preferences elicited purely from comparison is expected to yield explanations of valuation in microeconomics and psychology that hew closer to actual human behavior.

> *Wealth - any income that is at least one hundred dollars more a year than the income of one's wife's sister's husband.*
>
> H L Mencken

Details of the observation probability definition

Defining the observation probability $p(x|c)$ in terms of element-wise mismatches between the observation subset and the context subset of world possibilities requires us to maintain two indices over either type of subset. We use y^t to denote an indicator function on $\mathcal{X}$ encoding the possibilities observed as $o^{(t)}$,

$$y^t(x) = \sum_{i \in o^{(t)}} \delta(x - i).$$

Similarly, we index contexts with an indicator function z on $\mathcal{X}$, so that for context $c^{(t)}$,

$$z^t(x) = \sum_{i \in c^{(t)}} \delta(x - i).$$

Given this indexing, we can denote the element-wise mismatch probability as $p(\neg y_i^t | z_i^t)$. Since $p(x_i|c^{(t)}) = 1 - p(\neg y_i^t | z_i^t)$, we can use these element-wise probabilities to compute the likelihood of any particular observation $o^{(t)}$ as,

$$p(o^{(t)}|c^{(t)}) = 1 - p\left(\bigcup_i^{|o^{(t)}|} \{\neg y_i^t\} | \bigcup_i^{|c^{(t)}|} \{z_i^t\}\right) = 1 - \beta \sum_i^{|\mathcal{X}|} p(\neg y_i^t | z_i^t), \tag{2.6}$$

where β is a parameter controlling the magnitude of the penalty imposed for each mismatch observed.

To concretely instantiate our likelihood definition in (2.6), we define a specific mismatch probability,

$$p(\neg y_i^t | z_i^t) = \frac{1}{|\mathcal{X}|}\left((1 - z_i^t) y_i^t + (1 - y_i^t) z_i^t\right), \tag{2.7}$$

with $\beta = 1$ for all our demonstrations.

Proof of representational equivalence

To show that our measure of relative desirability R also completely represents preference information, it should be sufficient to show that, for any two possibilities $x_i, x_j \in \mathcal{X}$, and for any observation context c

$$x_i \succ x_j \Leftrightarrow R(x_i) > R(x_j). \tag{2.8}$$

Since the existence of preference reversals through context variation destroys the possibility of a stable preference relation, we begin by restricting our analysis to preferences that satisfy a **context consistency** requirement,

$$\exists c \in \mathcal{C}, s.t.\ x_i \succ x_j \Rightarrow x_i \succ x_j \forall c \in \mathcal{C}_{ij}, \{x_i, x_j\} \in \mathcal{C}_{ij} \subseteq \mathcal{C}. \tag{2.9}$$

This additional requirement makes the expression of preferences in the context-aware setting epistemologically equivalent to the standard characterization of binary preference, since an observer insensitive to context will simply find that $x_i \succ x_j$ whenever the two possibilities are observed together. To completely characterize a preference relation over $\mathcal{X}$, however, simply specifying consistent binary preferences is insufficient. Analogous to the regular concept of transitivity, we further assume the existence of **transitivity between contexts**, such that,

$$\text{if } x_i \succ x_j \text{ in } c_1 \text{ and } x_j \succ x_k \text{ in } c_2, \forall c \in \mathcal{C}, x_i \succ x_k, \tag{2.10}$$

thereby introducing a sense of preference order across observable contexts.

Now, consider that for any pair of possibilities $\{x_i, x_j\} \subseteq \mathcal{X}$, the set of observable contexts can be partitioned as,

$$\mathcal{C} = \mathcal{C}_{\setminus ij} \cup \mathcal{C}_{i\setminus j} \cup \mathcal{C}_{j\setminus i} \cup \mathcal{C}_{ij},$$

with the subscript indices indicating the possibilities from among $\{x_i, x_j\}$ considered feasible, i.e.$p(x|c) = 1$ within that context subset[7] . Let $\mathbb{C} = \{\mathcal{C}_{\setminus ij}, \mathcal{C}_{i\setminus j}, \mathcal{C}_{j\setminus i}, \mathcal{C}_{ij}\}$. Then, we can expand the desirability definition in Equation (2.2) to,

$$R(x) = \frac{\sum_i^{|\mathbb{C}|} \sum_c^{\mathcal{C}^{(i)}} p(r^{(t)}|x, c)p(x|c)p(c)}{\sum_i^{|\mathbb{C}|} \sum_c^{\mathcal{C}^{(i)}} p(x|c)p(c)}, \tag{2.11}$$

[7] Recall from the main text that the notation $\mathcal{C}_{i\setminus j}$ references the subset of all observed contexts that contain x_i but not x_j.

Using our definitions of $p(x|c)$ and $p(r|x,c)$ (see (2.7) and immediately contiguous text), it is straightforward to show that,

$$R(x_i) = \frac{k_i \sum_c^{\mathcal{C}_{i\setminus j}} P(c) + k_{ij} \sum_c^{\mathcal{C}_{ij}} P(c)}{\sum_c^{\mathcal{C}_{i\setminus j}} P(c) + \sum_c^{\mathcal{C}_{ij}} P(c)}, \quad R(x_j) = \frac{k_j \sum_c^{\mathcal{C}_{j\setminus i}} P(c) + k_{ji} \sum_c^{\mathcal{C}_{ij}} P(c)}{\sum_c^{\mathcal{C}_{j\setminus i}} P(c) + \sum_c^{\mathcal{C}_{ij}} P(c)}, \tag{2.12}$$

since all other contributions disappear due to corresponding entries in $p(x|c)$ being zero. Here, the single indexed k_i counts the number of times possibility x_i was considered the most desirable in contexts including x_i and excluding x_j; k_j being defined symmetrically. The double-indexed k_{ij} counts the number of times x_i is considered the most desirable possibility in contexts where x_j is also believed to be present. Again, k_{ji} is defined symmetrically.

From (2.12) it should be clear that, in general, differences in the sampling of contexts in an agent's history of observations, measured, for instance, as variations in the size of the context subsets $\mathbb{C}^{(i)}$ will render comparisons between desirability values undecidable[8] . Hence, to retain consistent preferences, we require an additional condition on the history of observation contexts that generate our relative desirability measure. Specifically, we assume,

$$\forall x_i, x_j \in \mathcal{X}, \lim_{t\to\infty} |\mathcal{C}_{i\setminus j}| = |\mathcal{C}_{j\setminus i}|, \tag{2.13}$$

reflecting the intuition that there be no informative reason underlying the partial observability of world possibilities, i.e., partial observability occurs via random subset selection from $\mathcal{X}$. Note that this assumption, by symmetry, also implies

$$\lim_{t\to\infty} p(x|\text{data}^{(t)}) = U(x), \tag{2.14}$$

$U(\cdot)$ representing the uniform distribution.

[8] To see why this must be the case, observe that for any two functions of homologous form to R such that $\frac{\alpha k_i + k_{ij}}{\alpha+1} = \frac{\beta k_j + k_{ji}}{\beta+1} + \theta$, with the k values fixed, it is always possible to find a new $\beta' = \beta\left(1 + \frac{\theta}{k_j} + \frac{\theta}{k_j(\beta+1)}\right) + 1$ that will reverse the inequality.

Given this, in the infinite data limit, we obtain

$$p(x|data) = \sum_{c}^{\mathcal{C}_{i\setminus j}} p(c) + \sum_{c}^{\mathcal{C}_{ij}} p(c) = \sum_{c}^{\mathcal{C}_{j\setminus i}} p(c) + \sum_{c}^{\mathcal{C}_{ij}} p(c) = U(x),$$
$$\Rightarrow \sum_{c}^{\mathcal{C}_{j\setminus i}} p(c) = \sum_{c}^{\mathcal{C}_{i\setminus j}} p(c),$$

obviating the necessity of further accounting for the denominators in (2.12).

It is now quite straightforward to demonstrate both directions of (2.5). First, assuming the left hand side of (2.5) immediately sets $k_{ji} = 0$. Further, using symmetry in context observability, k_i can now be interpreted as determining the number of times x_i dominates all other possibilities in $\mathcal{X} \setminus \{x_j\}$; k_j vice versa. By (2.10) x_i dominates all possibilities that x_j dominates, by (2.13) the number of observations over which either possibility can dominate is equal and by (2.14), in the limit of infinite decision samples, they will observe the same alternative possibilities, implying $k_i \geq k_j$. Since $k_{ij} > 0$[9] , we directly have,

$$k_i \sum_{c}^{\mathcal{C}_{i\setminus j}} p(c) + k_{ij} \sum_{c}^{\mathcal{C}_{ij}} p(c) \geq k_j \sum_{c}^{\mathcal{C}} p(c),$$
$$\Rightarrow R(x_i) > R(x_j).$$

Assuming the RHS of (2.8) to be true, adopting the selection rule $\max_x R(x)$ proves the converse. Hence, contingent on the three assumptions we have specified above, the relative desirability based decision framework encodes relative preference relations equivalently well as ordinal utility functions.

[9] Assuming the LHS of (2.5) forces k_{ij} to be at least 1.

Chapter 3

Cognitively efficient belief formation explains dynamics of human probability distortions

The behavior of human subjects in real decision-making tasks differs greatly from the normative expectations of expected value decision theory. We follow the prospect theory convention of defining *risk sensitivity* as deviation from linearity seen in the probability weighting function defined by [4]. Deviations from expected value predictions for risky choices have traditionally been explained by postulating the existence of varying levels of risk aversion in participating subjects. As the importance of risk preferences for decisions with important economic outcomes has become clearer, multiple studies [57, 58, 59, 60] have attempted to find causes for individual differences in risk aversion profiles. A common observation from these studies is that risk sensitivity appears to be positively correlated with cognitive ability, indicating both that subjects with greater cognitive ability are more likely to take risks in certainty-equivalence experiments [57, 58, 59]. It has also been observed that subjects with higher cognitive ability show more patience in inter-temporal choice settings [58, 59]. Furthermore, in a recent behavioral experiment,Zhang & Maloney [47] have shown a systematic increase in subjects' risk sensitivity based on increasing experience with choice options and the numerosity of choice samples.

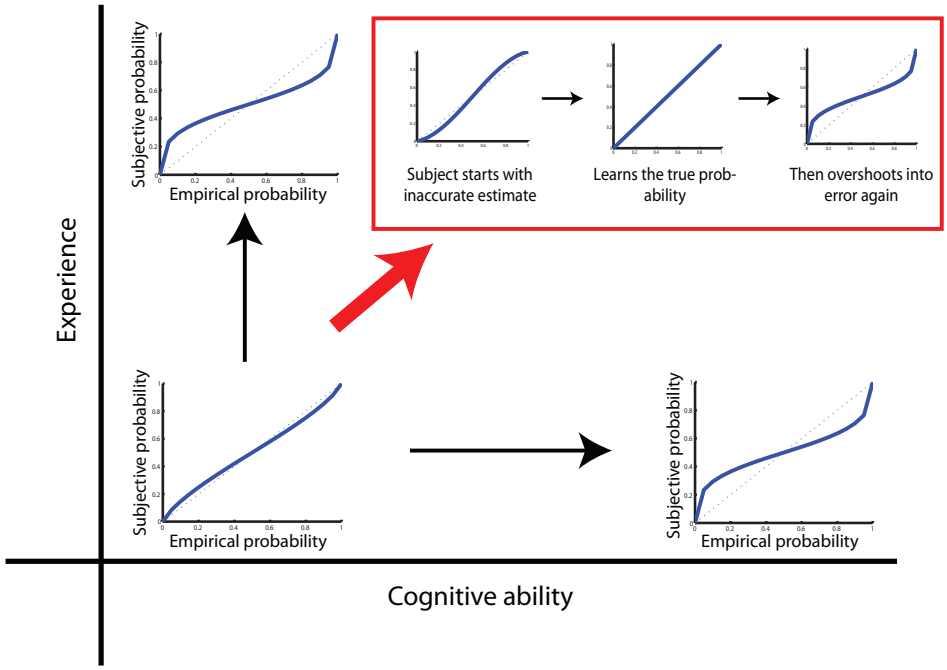


Figure 3.1: The non-linearity of the probability weighting function measuring risk sensitivity increases both with measures of cognitive ability, and with increase in experience with a particular risky prospect. This figure plots the expected variation in the weighting function along both axes. The origin of these effects is currently unknown, but expected to correspond to aspects of the cognitive processes that underlie risk sensitivity.

These empirical observations are tantalizing beyond simply uncovering individual and manipulable differences in risk sensitivity. One of the most promising avenues of research into economic behavior in recent times comes from the idea that the human brain is Bayes-optimal, that it learns what to do from observing the world according to Bayesian inference, and that it then acts according to the inductive theories of the world that it develops. As Griffiths & Tenenbaum [61] demonstrate, such a hypothesis leads to a natural explanation of a number of biases in human probability perception. However, data from the behavioral studies we cite above presents a different, and somewhat counter-intuitive picture. For instance, subjects in Zhang & Maloney's experiments begin with an underestimate of the probability of rare events, due to insufficient data sampling, as has also been observed in subjects making decisions from experience [62]. As their experience with a particular risky choice increases (see inset in Figure 3.1), subjects approach the true probability value, as a rational Bayesian agent would be expected to do. However, thereafter, the subjects tip over into over-estimating the probability of rare events, resulting in the familiar risk pattern observed by [4, 63] and others. Such deviation away from the true signal is maladaptive from the standpoint of Bayesian inference, and cannot be accomodated by existing theory. Similarly, [57] and [58] show that higher cognitive ability is associated with greater deviation away from normative choice in risky decisions. Across studies, a counter-intuitive pattern emerges, wherein *qualities that we generally associate with a better understanding of the world (i.e. intelligence, experience) result in a greater degree of irrational behavior.*

In this work, we show how using cognitive modeling to address what has heretofore been strictly a classical economics problem, yields a surprisingly simple potential explanation for these paradoxical results. The very idea of risk, in the strict economic sense, emerges from deviations from perfectly utilitarian rationality expected from human subjects in behavioral experiments. Thus, what is measured as risk, is, in fact, simply the generation of beliefs about whether to prefer a risky choice or not, based on iterated experience. Rather than use a static view of utilities as being embedded in the environment, as is prescribed in neo-classical economics treatments, an intuitive case can be made for studying the problem of preference development for risky options as one of learning useful beliefs from experience. Our principal contribution in this paper is to show that a reformulation of the economic risk analysis problem as one of

cognitive belief formation provides a novel mechanistic explanation of the emergence of risk sensitivity. In the remainder of this paper, we describe how such an explanation emerges from a dynamic cognitive model of belief formation; explaining the data relating risk sensitivity with cognitive ability and experience endogenously and leading to novel testable behavioral predictions, a deeper operationalization of related economic and behavioral concepts, and clear policy implications.

3.1 Dynamic belief formation

Embodied decision-making agents are expected to form preferences about options they encounter dynamically in ways that economic choice models fail to describe. Economics choice models [cite discrete choice models], however, measure subjects' evaluations from revealed preferences, eliding the role of memory in the formation of these preferences. While such choice models are useful from the standpoint of practical economics, wherein all we need is a manipulable *description* of subject behavior, they are inapplicable as *theories* of belief *formation*, and can explain neither the basic distortions of probability observed by [4, 63], nor the systematic changes in these distortions observed in the [57] and [47] studies.

Belief formation implies that beliefs about an experience must be formed dynamically based on past experiences. The belief formation process, hence, is inextricably linked with the process of memory recall. A number of cognitive architectures elucidating the process of memory recall [64, 65, 66] using a production-system model of memory [67] have been developed in the past two decades. The production system memory model assumes that agents have long-term stable memories of past experiences, from which they draw a subset of these experiences into a more limited and *dynamic* working memory, and make future decisions based on this subset of experiences. Under various further assumptions about the mathematical representations of these experiences, semantic, episodic and reinforcement learning-based recall procedures can be simulated [68] using such models. For our purpose, it is sufficient to assume the most general production system model - one which requires only the existence of a working

memory intermediating belief formation from past experience. That is,

$$p(x) = \sum_{m \in \mathcal{M}} p(x|m)p(m), \tag{3.1}$$

where $x \in \mathcal{X}$ are the choices available to the agent, and $m \in \mathcal{M}$ are experiences present in memory. In a production system sense, the probability distribution $p(m)$ encodes the likelihood of recalling the memory of experience m. For our present purpose, we assume that agents have direct access to $p(x|m)$ at sequential time instances, since inferring desirability in situ is a different problem from the one we are addressing.

The recall distribution $p(m)$ can be generated using a number of different memory models. Primed sampling methods assume a flat $p(m)$. Assuming that experiences *similar* to the present experience have greater $p(m)$ leads to the ACT-R [64] memory model. Reinforcement learning methods [8] assume that more recent experiences are more likely to be recalled. We [16] have recently proposed that belief formation results from agents attempting to minimize cognitive effort while making decisions. The resultant model of belief formation is seen to behave in accordance with prospect theory predictions, making it a suitable candidate for investigating the relationship of risk sensitivity to belief formation.

The cognitively efficient belief formation hypothesis postulates that humans try to minimize their metabolic costs during belief formation by recalling as few past experiences as feasible to yield beliefs that satisfactorily predict the environment. Efficient recall involves determining which memory samples are likely to be most informative. In general, past beliefs can either support the agent's current plan of action, or oppose it. Therefore, informative beliefs will either be completely congruent with the agent's latest belief (supporting the current policy), or strongly violate it. This intuition can be quantified in terms of the *surprise* experienced by an agent operating with the current belief $p(x)$ in comparison with a stored belief $p(x|m)$ can be quantified by an information divergence [69],

$$R(p(x), p(x|m)) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p(x|m)}. \tag{3.2}$$

The informativeness of a stored belief $p(x|m)$ can be measured as the deviation from the average surprise $\overline{R}$ experienced by the agent in its event history,

$$A(m) = |R(p(x), p(x|m)) - \overline{R}|, \tag{3.3}$$

and is expected to be inversely related with the cognitive or metabolic cost of recalling this particular belief and hence, with the memory distribution p(m)as well. In practice, we use a softmax mapping to transform the inverse of $A(m)$ to $p(m)$. Given this generative process, a cognitively efficient agent will recall a limited subset $\mathcal{M}'$ of all possible experiences that permits reconstruction of what the agent can expect to be a sufficiently predictive belief.

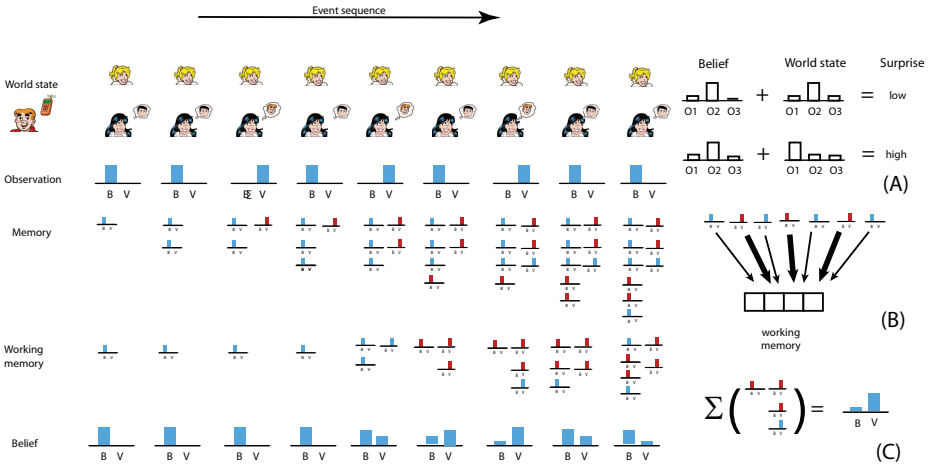


Figure 3.2: Illustration of the cognitively efficient belief formation model on a simple decision task in which Archie must decide who to ask out on a date, Betty or Veronica. The model updates beliefs by combining new observations with a distribution on beliefs, where the belief distribution is constructed bringing samples of the memories past observations into working memory. A cognitively efficient agent recalls a limited subset of beliefs about past observations of the world state to inform its current belief. At the first event, previous experience in memory suggests Betty will never say no, but Veronica is likely to. Cognitive efficiency is obtained by recalling the smallest number of past experiences necessary to synthesize a usable belief. Beliefs about past observations are prioritized for recall based on the amount of information they are expected to contain. **(A)** Beliefs that completely support or contradict observations about the world state are informative and **(B)** are prioritized during belief recall. The average of all beliefs collected in working memory at a particular decision instance **(C)** gives us the agent's current belief.

A deeper comparison of the relative merits of our dynamic decision model in comparison with existing proposals lies outside the scope of the present work. For our present purposes, it is sufficient to note that, unlike expected utility models favored in neoclassical economics, our approach leads to a dynamic model of belief formation

that involves a functional relationship between belief uncertainty and working memory utilization. However, it is well-known that working memory in biological agents is limited in size. Therefore, it is evident that working memory size constraints must affect belief formation in agents that behave according to our theory. A substantial body of literature implicates working memory as being strongly correlated with measures of cognitive ability [70]. Fukuda et al. [71] have recently shown further, that it is the size of working memory that primarily influences this correlation. If we assume working memory size as a measure of cognitive ability, then the structure of our belief formation theory allows us to simulate experiments relating both cognitive ability and experience with risk preference behavior.

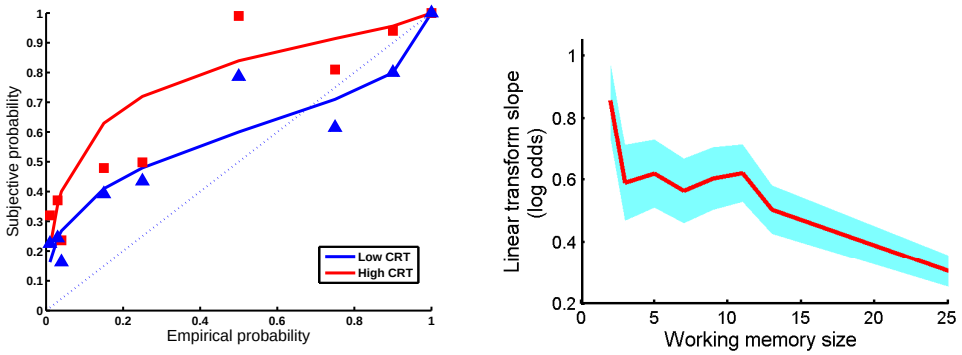
3.2 Results

Our model of belief formation directly predicts choice probabilites for presented options, which may then be further transformed into other economic measures of relative preference. We conducted simulation-based experiments on a binary outcome space, controlling the sequence of trials simulated agents saw to appears as if resulting from a pairwise comparison between a moderate-value safe option (M) and the actual payoff of a risky gamble between a high value option (H) and a low value option (L), such that $H > M > L$. Our theory of belief formation results in an endogenous replication of results from [57] showing increased risk seeking for low probability gains in high cognitive ability subjects, results from [47] showing increased non-linearity of probability weighting with increasing task experience, and results from [58] and [59] showing greater patience in intertemporal choice and lesser risk aversion in risky choice for subjects with greater cognitive ability.

Cognitive ability effects on risk sensitivity. Since working memory size is strongly correlated with fluid intelligence, manipulating this parameter of our belief formation model can be considered a manipulation of cognitive ability of the simulated agent. Hence, by operationalizing agents with dynamic preference formation computationally in a standard risky vs safe prospect selection setup (see SI for details), we obtain predictions about how such agents will behave under a range of payoff probabilities and gain/loss framings. Comparing such behavior across a population of agents

heterogeneous in terms of memory size limitations leads to computational predictions connecting cognitive ability with risk sensitivity.

Our simulation setup allowed us to observe the behavior of agents when offered multiple p-prospects and test the relationship between working memory size and risk sensitivity across the entire range of prospect probabilities. To this end, we replicated the methodology of the previous simulation for 21 different values of p, evenly spaced between 0% and 100% chances of winning $10. Compiling the average 5-back choice probability for the risky option at the end of a 200 time step training history for each agent, we directly obtained decision weights for all 21 gambles, giving us an empirical estimate of the probability weighting function describing each agent's choices. Comparing the shapes of the average probability weighting function obtained for each working memory size sub-population, note (Figure 3.3(b)) that agents with larger working memories have probability weighting functions that show greater deviations from risk-neutral behavior across the range of payoff probabilities.



(a) A log odds function with a smaller slope parameter better fits probability decision weights that describe the choice behavior of high cognitive ability subjects in the experiments in [57].

(b) Slope of linear transformation decreases with working memory size for cohorts of simulated cognitively efficient agents.

Figure 3.3: This plot verifies a negative correlation between working memory size and risk sensitivity in both human data and model predictions. The mean and the standard deviation for the slope parameter are computed using Monte Carlo fits given the final choice probability of each agent.

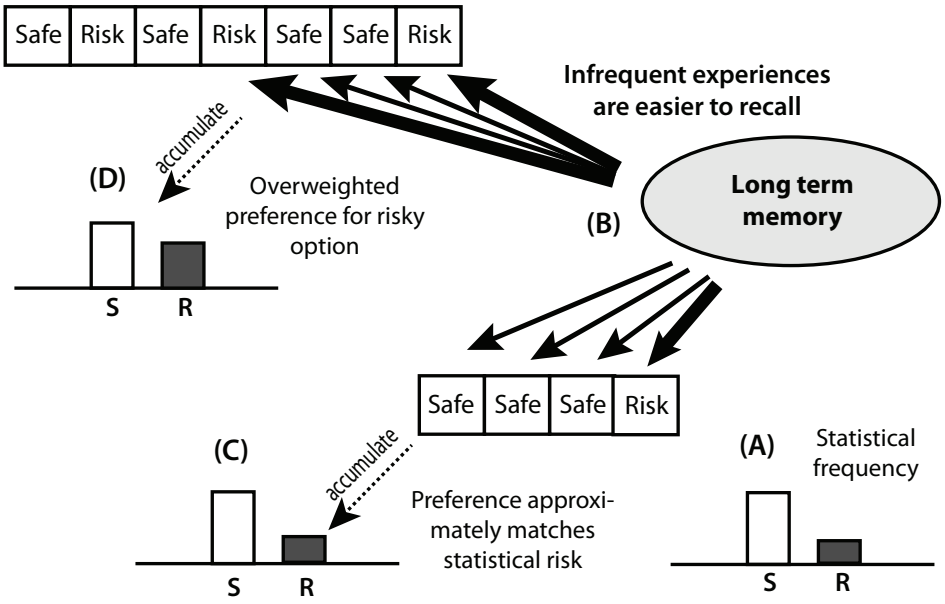


Figure 3.4: How risk sensitivity arises: cognitively efficient preference formation prioritizes recall of exceptionally typical or exceptionally atypical beliefs, which largely correspond in this case to selecting safe and risky options respectively. This biases the agent to expect more extreme outcomes than warranted purely by the probabilities of outcomes. Agents with smaller working memories sample overweighted risky options less frequently than those with larger working memories, purely as a matter of requiring fewer random samples to populate working memory. Once a risky option has been sampled, it dominates preference construction, since only preferences recalled into active memory are averaged.

What causes this relationship between risky preference formation and working memory size to emerge? Since agents recall beliefs to the extent necessary for accurate predictions, constrained by working memory size, agents with larger working memories have a greater chance of recalling both high and low regret choice instances, leading to an overweighting of low probability outcomes. On the other hand, agents with small working memories end up recalling only low regret choices, which makes them prefer to gamble pN times and prefer safety (1 - p)N times over N trials with a p prospect, resulting in risk neutral approximate probability matching. Thus, the relatively straightforward use of a dynamic model of belief formation in risk preference simulations leads to the apparently counter-intuitive, but empirically suported, prediction that subjects with lower cognitive ability will behave in a manner closer to risk neutrality than subjects with greater cognitive ability.

Experience effects on risk sensitivity Zhang & Maloney [47] document an increase in belief distortion in human subjects with increasing experience with gambles. Participants in their experiment were asked to estimate the relative frequency of either black or white dots against a backdrop containing dots of the complementary color. By varying the relative proportion of these colors, they were able to test subjects' sense of relative frequency across the entire range of probabilities. Interpreting subjective judgments of normalized relative frequency as choice probabilities, our model directly predicts the subjective decision weighting elicited by Zhang & Maloney. We simulated our model with an identical experimental setup, leading to the elicitation of choice probabilities commensurate with their basic finding. Figure 3.2 shows that our model's behavior closely matches that of human subjects both qualitatively and quantitatively across 8 experimental trials. We note with particular interest that the close quantitative relationship between the model's predictions and human data emerges with no statistical fitting, suggesting that the degree of probability distortion seen in such experiments might emerge from basic structural aspects of the cognitive architecture.

We also observe that individual agents in our simulations begin with $\lambda > 1$, a condition that [62] describe as natural to decisions from experience, and then, with further experience, shift towards behavior better described by $\lambda < 1$, viz. behavior expected in decisions from description [48]. This natural shift between these two regimes of probability distortion argues in favor of the contention that they may not be conceptually

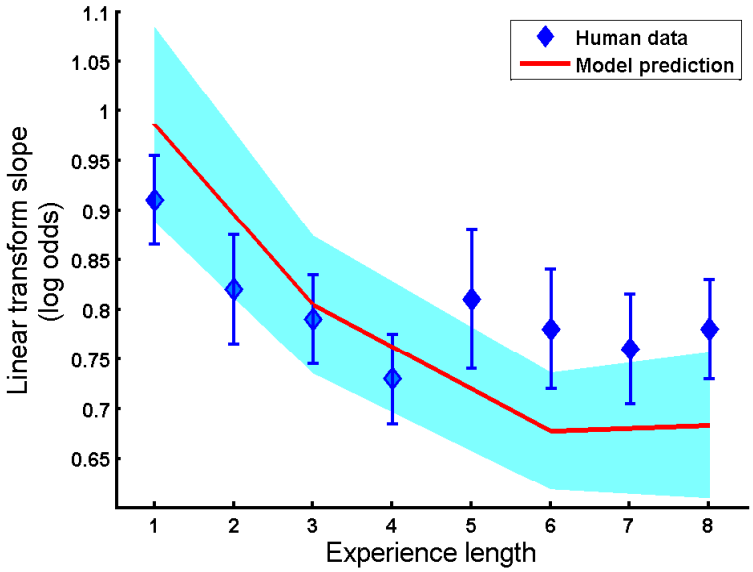
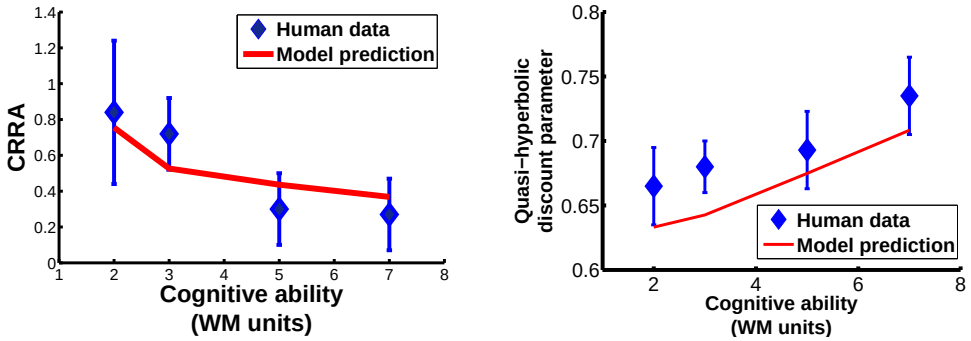


Figure 3.5: Slope of linear transformation measuring deviation of model behavior from risk-neutrality decreases with increasing experience with the particular risky choice for population cohorts of equal size (N = 20). Human data is obtained from Figure 8 in [47]. Mean and the standard deviation for the slope parameter are computed via Monte Carlo fitting using final choice probabilities of all agents.

as disparate as previous studies (see e.g. [72] for a recent review) have assumed, and in fact, are reconciliable simply through giving agents making decisions from experience a larger learning sample.

Cognitive ability effects on risk aversion. Both [58] and [59] report a decrease in risk aversion for 50/50 gambles. The former study reports this in the form of a reduced coefficient of relative risk aversion for higher cognitive ability populations, whereas the latter shows an increasing trend in the certainty equivalent that causes a switch from risky to safe option for subjects with greater cognitive ability. [58] use the standard RAPM measure of cognitive ability, whereas [59] use a unique cognitive measure designed for use in verbal interview settings.



(a) Model predicts decrease in risk aversion for choice under uncertainty with increasing cognitive ability, commensurate with observations in human subjects [58]

(b) Model predicts increase in patience in intertemporal choice with increasing cognitive ability, commensurate with observations in human subjects [58]

Figure 3.6: Comparison of model predictions with human data collected by Burks et al [58]. Choice probability can be mapped to CRRA assuming that human agents in certainty switching experiments have perfect probability perception and isoelastic utilities; assumptions identical with the original study. Discounting factor computations are also equivalent to those performed in the original study. WM units of cognitive ability are estimated from [71] and remain fixed across all result replications.

As we demonstrate in SI Methods, it is possible to estimate relative risk aversion from choice data by replacing prospect theory assumptions with expected utility assumptions in the certainty equivalence tasks used in the human studies in [58] and [59], accurate to within ambiguity about the crossover point in the probability weighting function. For each simulated agent's learned choice probability for the risky option, we computed

the implied coefficient of relative risk aversion (CRRA), as described in SI Methods. Figure 3.6(a) plots the mean CRRA for simulated population cohorts with different working memory sizes. The results show a clear inverse relationship between working memory size and CRRA, with the difference between endpoints statistically significant ($p < 0.005$), in concord with the predictions of [58]. This finding supports the hypothesis that working memory size is directly correlated with increasing risk sensitivity. Similarly, as shown in Figure 3.6(b), simulated agents with greater cognitive ability showed greater patience in inter-temporal choice experiments, as empirically observed in [58].

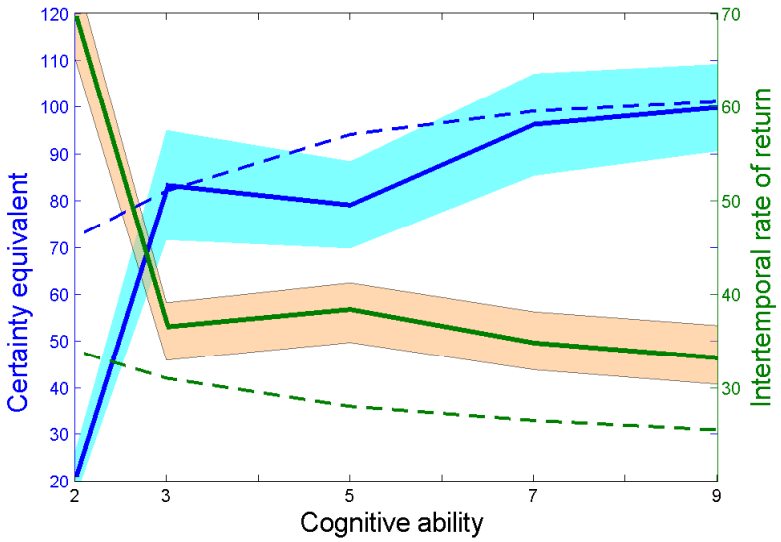


Figure 3.7: Model replicates [59] results on risk aversion and intertemporal patience. Dashed lines show smoothed values from human data, taken from the original paper. Certainty equivalent points on the graph are proportional to the model's final choice probability after learning on a sequence of gambles with a fixed payoff probability. Intertemporal rate of return can be computed from the quasi-hyperbolic discounting factor, as described in SI Methods. Payoff probability was varied and $p = 0.05$ gave the best fit. All other probability values retrieve the same trend, but at different scales. WM units of cognitive ability are estimated from [71] and remain fixed across all result replications.

The main results from Dohmen et al [59] are also replicated using a similar methodology. Dohmen et al report an increase in assessed certainty equivalents in a task design identical to that used in [58], as well as a reduction in an acceptable rate of return in

an intertemporal choice task using the same design. We replicate these experiments using simulated agents (see SI Methods for details) and plot results in Figure 3.2. The predictions of our model are entirely in line with the conclusions of [59].

Interestingly, we find close quantitative fits to both the Burks and Dohmen studies using an identical mapping from nominal working memory sizes used in our simulations to cognitive ability percentiles assessed in the two studies (see SI Methods for a fuller discussion of this point). This invariance across two different measures of cognitive ability reassures us that the nominal working memory measurement of cognitive ability is not without value.

3.3 Discussion

General implications: Our findings support the basic conclusion that agents with greater cognitive ability will be relatively less risk averse in selecting risky options with low gain probabilities, mirroring the empirical results of [57, 58, 59] and that this difference is sustained independent of whether the objective expected value of the gamble is lower than the value of the safe option, as shown in [57]. Our simulation further predicts analogous behavior in the other three quadrants of the prospect theory risk pattern, which leads to testable predictions summarized in Figure 4.8(a). We further find that agents with greater cognitive ability wil delay gratification in intertemporal risk trials, supporting concordant observations in human subjects in [58].

Our proposed mechanism explaining the origins of risk appetite presents an interesting contrast with existing theories explaining the effect of intelligence on risk preferences. Burks et al's signal processing hypothesis [58], assumes that utility perception involves processing a noisy signal, and that the amount of noise in the utility signal is inversely proportional to cognitive ability. This theory predicts behavior closer to risk neutral for subjects with greater cognitive ability, and hence fails to explain increased risk seeking beyond rational utility expectations in subjects with higher cognitive ability, as shown in [57]. We, on the other hand, find that it is, in fact, more intelligent agents who are capable of encountering greater noise, by virtue of a greater 'bandwidth' for recollection and assimilation of experiences.

Dohmen et al [59] suggest two possible mechanisms that could potentially explain

this relationship: one assumes that subjects with lesser cognitive ability bracket choices [73] more narrowly, resulting in increased risk aversion, the other assumes that a deliberative PFC-centered utility computation is dominated by an affective, myopic emotional response mechanism in subjects with lower cognitive ability. The 'two-system' proposal, while it has some neurobiological substantiation [74], suffers from the same lacuna as the utility signal processing hypothesis - it fails to account for risk-seeking behavior for low probability gains. The choice bracketing argument, while potentially accurate, leaves open the question of how narrower choice bracketing leads to lower risk sensitivity. Our proposed mechanism appears to dovetail well with the choice bracketing hypothesis, since the latter essentially assumes that subjects with lower cognitive ability assimilate fewer past choices while making their current decision. Thus, in addition to demonstrating the emergence of risk appetite, the cognitive mechanism we have proposed here also presents as a reasonable operationalization of the idea of choice bracketing in the context of sequential choice problems. Since choice bracketing has not heretofore been mechanistically operationalized, our current proposal serves as a possible mechanistic explanation for its origin just as well.

While *sturm und drang* behavior (buying lottery tickets and obsessing about the end of the world), as predicted for greater cognitive ability in this study, might appear maladaptive and contrary to general assessments of intelligence at first sight, it must be remembered that on evolutionary time scales, for foraging species like ours, actual world environments are non-stationary, with exploratory behavior likely to result in fitness jackpots and/or abysses. Thus, behavior that was extremely sensitive to risk, both in terms of taking long shot risks and in insuring against unlikely calamities, are likely to have been conferred selection benefits.

Finally, it has not escaped our attention that our mechanism for the emergence of probability distortions also doubles as a mechanistic account for the probabilistic distortions characterizing prospect theory. As [75] point out, 'establishing a neural and evolutionary basis of prospect theory could provide an illustrative example of how the foundation for principles guiding social science might be usefully shifted from relying largely on logic, to respecting biological implementation.' We believe that our present effort, while not seeking to explain framing effects and loss aversion, makes a contribution in this very important direction.

Related results: Dohmen et al. [59] point to data from two other studies whose conclusions can be reinterpreted in light of our results. In a study conducted by Shiv and Fedorikhin [76] it was seen that subjects who are required to keep in mind a seven-digit number while selecting among food options are more likely to choose an unhealthy snack over a less enjoyable but healthful option than matched controls who are not required to remember a number. We interpret this as a reduction in the intertemporal discounting rate - preferring short-term gain (taste) over longer-term gain (health) - caused by an artificial reduction in working memory size through cognitive loading, which concurs entirely both with the predictions from our simulation and the cognitive mechanism we have proposed to explain its operation. Benjamin, Brown, and Shapiro [77] also describe qualitatively similar results, where inducing a cognitive load leads to more impatient and more risk-averse decisions, again in concord with our predictions. The data from these two studies, while not directly relevant to our simulation, are important because unlike conceivably static genetically induced differences in cognitive ability, cognitive load experiments actively manipulate working memory size. Therefore, differences in risk appetite generated actively through such treatments indicate the involvement of an active causal mechanism for the generation of risk preferences, thereby opposing nativist explanations grounded in static neurobiological factors, e.g. signal processing noise, greater affective response etc.

The mechanism we propose for the emergence of risk-sensitivity also provides an interpretation for the correlation between Prelec's non-linearity constant [78] inferred from subject behavior in test tasks and activation in the anterior cingular cortex observed by [79]. Since the ACC is well-known to be strongly implicated in conflict monitoring and emotive response, we hypothesize that the greater activation measured for more non-linear weightings reflects the ACCs tracking of the conflict necessitated in combining divergent beliefs in working memory during preference formation.

Finally, to the extent that this research suggests that quick assessments of risk are likely closer to statistical truth than more deliberate consideration, particularly for subjects with high cognitive ability, it also presents potentially supporting evidence for the existence of 'thin-slicing' phenomena in decision-making, where it is seen that the accuracy of the predictions of proficient forecasters reduces with the amount of time they are given to make predictions [80].

Limitations of our memory model: Unlike in economics, psychological decision-making models have long understood the importance of memory recall in the formation of beliefs indicating preferences for different options. Our own model of belief formation resembles the ACT-R family [64], differing primarily only in the manner of memory recall. Specifically, the salience of past experiences is operationalized in an information theoretically efficient manner, as has been proposed in [81], based on the degree of predictive surprise experienced when comparing beliefs corresponding to the past experience with beliefs corresponding to the present experience. This assumption dovetails well with recent neuroanatomical evidence implicating the role of norepinephrine as a neural correlate for predictive suprise (Dayan, 2006). The dyadic sensitivity both to extremely surprising and extremely unsurprising events arises out of information theoretic considerations (such experiences contain the most information) and finds empirical support in recent evidence presented by [82]. We note that our theory of cognitive efficiency instantiates the first production-system memory model that can replicate the prospect theory probability distortions, and so, is the only one that can be used in the current experimental setup.

While it may be feasible to estimate the effect of memory size limitation on risk preferences using other memory models, we believe that the basic intuition of belief formation as an importance sampling over the set of past experiences must necessarily emerge as the causal factor leading to the over-weighted availability of rare events. Once rare events are over-weighted, the rest of our economic results directly follow. Hence, notwithstanding the details of the cognitive architecture used for such alternative models, the underlying theory for the emergence of risk appetites in binary lotteries, and it's predictions on the impact of working memory size, as illustrated in Figure 3.4 should remain unchanged.

Policy implications and future research: The connection between cognitive ability and risk preference is one of the first examples of actionable evidence of individual differences in economic behavior to have come to light, and has important policy implications. For example, as Boyle et al [60] point out, a strong negative correlation between cognitive *performance* and risk aversion persists into old age, a period of life where senior citizens make economic and health-care decisions that disproportionately

affect the fiscal condition of the modern welfare state. Given that cognitive function declines gradually with advancing age, a causal understanding of the relationship between cognitive ability and risk sensitivity could generate new insights for assisting senior citizens showing cognitive decline make better decisions.

Our model predicts that subjects will become increasingly risk averse to losses with low probabilities with increasing experience with a binary choice. In domains where qualitative decisions with precisely this structure are made (e.g. end-of-life care, default risk assessment) our model would suggest that health providers/risk adjusters with an intermediate degree of experience will make the most accurate predictions. Such a prediction, if true, should result in a significant re-evaluation of organizational practices in these domains. On smaller time-scales, our model also motivates time-boxing/Pomodoro style multi-tasking techniques that explicitly limit exposure to particular problems as a means of promoting efficiency.

Methods

General cognitively efficient belief formation The cognitively efficient belief formation hypothesis postulates that humans are essentially searching for minimal-cost theories about how to choose high value options, where the cost is measured in terms of the complexity of encoding and storing the information needed to reliably make these decisions. We term this cost *cognitive processing cost*, which is equivalent to the cost of accessing past beliefs in the agent's memory. Informally, to make a sequence of decisions, the agent cycles between forming relative preferences about the relative worth of options by accessing past experience, making choices, experiencing outcomes and updating these beliefs to minimize processing costs for future decisions. More formally, the agent tries to minimize its cognitive processing cost T while maintaining a high level of predictive confidence C in the quality of its choices. The self-motivated learning objective is to minimize a function of the form:

$$\begin{aligned} \underset{\mathbf{x}}{\operatorname{argmin}} \quad & T \\ C_{\text{new}} \quad & \geq C_{\text{old}}. \end{aligned} \tag{3.4}$$

where T and C are defined below in terms of relative preferences. The basic idea

is that the agent updates its belief about the relative quality of items or users, but also stores a subset of past beliefs to track its own predictive confidence and to enlarge its experience when novel items are encountered. We represent an agent's belief about the relative worth of options at an event t $\mathbf{x}_t(s)$ as a probability distribution across the items $s \in \mathcal{S}$ available to it. Given environmental feedback, this distribution $\mathbf{x}_a$ can change or compared to other agent's beliefs, generating surprise. To track its predictive confidence, the agent computes surprise with respect to stored beliefs, which gives a measure of novelty and the ability to assess how divergent the current belief is from its past, or other's beliefs. Surprise experienced by an agent operating with relative preference $\mathbf{x}_a$ in comparison with relative preference $\mathbf{x}_b$ can be quantified by an information divergence [69],

$$R(\mathbf{x}_a, \mathbf{x}_b) = \sum_{j=1}^{n_a} \mathbf{x}_a^j(s) \log \frac{\mathbf{x}_a^j(s)}{\mathbf{x}_b^j(s)}. \tag{3.5}$$

With each update, the agent experiences some change in belief, but most changes are uninteresting. We propose that the agent tags changes that deviate strongly from the expected level of surprise, which we term predictive exceptionality. We measure predictive exceptionality of a stored relative preference $\mathbf{x}_{\text{old}}$ as the deviation from the average surprise experienced by the agent $\overline{R}$:

$$A(\mathbf{x}_{\text{old}}) = |R(\mathbf{x}, \mathbf{x}_{\text{old}}) - \overline{R}|, \tag{3.6}$$

where $\mathbf{x}$ is the agent's current relative preference. Finally we assume that the agent compresses its memory by discarding most stored relative preferences, but saving the typical and those that have high predictive exceptionality. We can capture this idea by constructing a cost T on storing and retrieving a subset $\mathcal{M}'$ out of the set $\mathcal{M}$ of all past relative preferences by summing the inverse of the predictive exceptionality score:

$$T = \sum_{\mathbf{x}_i \in \mathcal{M}'} A^{-1}(\mathbf{x}_i), \tag{3.7}$$

Finally, the model tracks its *confidence* $C : \mathbf{x} \to [0, 1]$ in its relative preferences, via a measure that grows when the relative preferences have low uncertainty and low surprise:

$$C = \frac{1}{C_{\max}} \frac{\log |\mathbf{x}| - H(\mathbf{x})}{\sum_{\text{memory}} R(\mathbf{x}, \mathbf{x}_{\text{old}})}, \tag{3.8}$$

Our model uses a simple belief update mechanism to incorporate environmental feedback, and it solves the optimization problem specified in Eqn 6.2 to decide which past beliefs to save and which to discard.

Prospect simulation We use a standard risky prospect setup for our experiments. Subjects (in our case, simulation trials) are presented with a choice between a safe option that pays \$2 and a risky option that pays \$10 with an unknown probability. A risky prospect that, statistically speaking, pays off p% of the time is henceforth called a p-prospect. Since the choice model we use directly gives us relative preferences, we consider the relative preference assigned to the risky option by an agent that has had sufficient experience with a p-prospect to correspond to the agent's subjective decision weight for that prospect. Sufficiency is measured in statistical terms through convergence of the decision weight to a stable value.

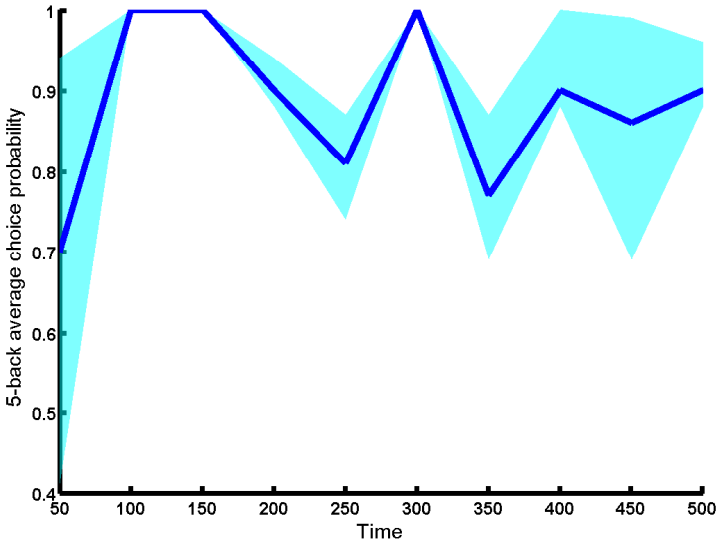


Figure 3.8: Cognitively efficient agents express relative preferences for different outcomes based on a history of similar past choices experienced. In our experiment, the relative preference ascribed to the risky prospect after a prolonged period of exposure to gambles with the same level of risk gives us the agent's learned subjective choice probability for the gamble. This plot shows that the choice probability for a risky gamble does, in fact, converge to a stable small range of values given a sufficiently large training history.

Figure 3.3 shows typical convergent behavior of agents' relative preferences given

repeated (> 200) encounters with a p-prospect. We further average the final decision weight over its last five elicitations to compute our simulated agents' decision weights for learned prospects.

Measuring risk sensitivity in prospect theory By varying p between 0 and 100 for different histories of prospects, we elicit decision weights from our choice model, as described above, and construct a probability weighting function for each agent's history of choices. We use least-squares fitting in log-log space to fit the decision weights we obtain for different agents' choice histories across p-prospects, $p \in [0, 100]$ to the linear log-odds weighting function,

$$\log \frac{w(p)}{1 - w(p)} = \lambda \log \frac{p}{1 - p} + (1 - \lambda) \log \frac{p_o}{1 - p_o}, \tag{3.9}$$

where λ is a slope parameter that takes the value of 1 for risk neutral behavior and incrementally deviates from this value for greater risk sensitivity, whether aversive or seeking. p_o represents a cross-over point where subjective and objective probability judgments coincide in the data sample.

Prospect theory interpretation of data from [57] We assume that the percentage of subjects that prefer the risky option in the [57] experiment can be assumed to be the choice probability c of the risky option r instead of the safe choice s for the respective cognitive ability-wise stratified population cohorts. We assume that subjects' choice probabilities c depend on their perceived option utilities following the Luce-Shepherd choice rule,

$$p(c_i) = \frac{\exp(U_i)}{\sum_i \exp(U_i)}, \tag{3.10}$$

and that the agents utility function, U, is isoelastic. A prospect theory-based formulation of the choice rule would imply,

$$\begin{aligned} \frac{\exp(\mathbf{w}(p) \times U(r))}{\exp(U(s))} &= \frac{c}{1 - c}, \\ \Rightarrow w(p) &= \frac{1}{U(r)} \left[U(s) + \log \frac{c}{1 - c} \right], \end{aligned}$$

where $w(p)$ is the subjective probability obtained from the prospect theory weighting function for the empirical probability p. Assuming the CRRA of the utility function to be approximately 1, based on summary empirical evidence from human data [83], we obtain probability weights for all the empirical probabilities specified in [57] for gains

and fit the best log odds weighting function to both the low and high cohort using least squares fitting in log-log space.

Measuring risk aversion Previous studies measuring effects of cognitive ability on risk attitudes have restricted themselves to a single prospect condition, and used expected utility methodologies for estimating risk preferences [58, 59]. Since we measure choice probability directly in our simulation, we do not follow the traditional expected utility methodology of varying the offered safe reward until the subject consistently prefers it over the risky option. Therefore, we have no direct measurement of subjective utility and consequent measures of risk aversion, a la [58]. We therefore elicit it from our data using a different methodology

For a binary choice between a safe option s and risky option r, an expected utility formulation of Luce's choice rule would indicate,

$$\frac{\exp(\mathbf{p} \times U(r))}{\exp(U(s))} = \frac{c}{1-c},$$
$$\Rightarrow U(r) = \frac{U(s)}{p} \log \frac{c}{1-c},$$

where $U(\cdot)$ is the underlying utility function mapping monetary value to hedonic reward, and c is the observed choice probability for the risky option. Since the safe option remains fixed across trials, we further assume $U(s) = k \times p$ to normalize the observed utility function for future calculations.

As in [58], we further assume that the agents utility function is isoelastic, i.e.

$$U(r) = \frac{r^{1-\rho}}{1-\rho}, \tag{3.11}$$

where ρ is the coefficient of relative risk aversion (CRRA) for a gamble. Since r, the nominal expected value of the risky lottery, is fixed in all experiment trials in [58] and U can be computed given choice probabilities from our simulation, we can compute ρ methodologically identically with the CRRA computation for human subjects in [58] given our model's simulated choice probabilities by numerically solving

$$r^{1-\rho} - U(r)(1-\rho) = 0. \tag{3.12}$$

Dohmen et al [59] presented subjects with a choice between a 50/50 lottery with fixed low and high values and a safe option that was initialized at the same value as

the low option in the lottery and incremented in constant increments until the subject switched their preference from the risky option to the safe one. This procedure allowed them to obtain a measurement of subjects' certainty equivalent value[1] . Since the certainty equivalent $CE = w(p) \times U(r)$, and $U(r)$ does not change across different trials in the [59] experiment design, $w(p) \propto CE$. Hence, CE values can be equated with agent's simulated choice probabilities (scaled to match measured variance).

Simulating intertemporal choice and measuring patience To simulate intertemporal risks for our agents, we set up a sequential binary choice problem where simulated agents have a choice between selecting a safe option now (at any time instant, with some fixed small probability) or a risky option at some future time, with the future time drawn from a uniform distribution on all possible future times. The later option was always set as more rewarding (larger) at the time of the gamble being offered, but could be reduced directly to zero according to a set hazard rate specific to the agent's environment. Agents with a history of experience with a fixed number of these gambles (N = 30) in their past choice history are then asked to select between safe and risky options in a setting identical to the sequential choice experiment above, with the average 5-back probability computed every 25 time steps from the 100^{th} time step to the end of testing (250 time steps), thus creating 6 time blocks. We average the choice probabilities obtained at the end of each time block across 200 different agents to average out history effects. The results presented in the paper hold for a wide range of hazard rates, with low hazard rates generally corresponding to lower absolute discounting rates.

Assuming an agent selects a later option (reward at time $t' = t$) with choice probability c at time $t' = 0$, it follows from the expected utility hypothesis that the future discounted expected utility of this option at the present time is

$$U^{(0)}(r) = c \times U^{(t)}(r). \tag{3.13}$$

At the same time, fitting a quasi-hyperbolic discounting model yields,

$$U^{(0)}(r) = \delta^t \times U^{(t)}(r). \tag{3.14}$$

[1] However, while the expressed winning probability for each lottery was 50%, subjects were further informed that they would receive monetary payments as determined by their performance in the task only 1/7 of the time. We suspect that this further specification in their experiment design shifts subjects from a choice frame where they are evaluating evenly matched probabilities as evenly matched to one where 'wins' are rare, thereby shifting risk preference into the risk-seeking lobe of the PT probability weighting function.

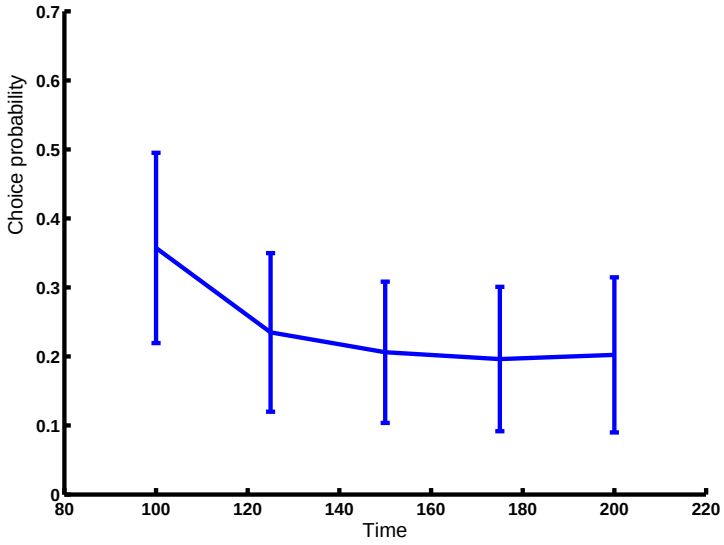


Figure 3.9: Plot of average 5-back choice probability for later (risky) options for a population of 200 agents with fixed working memory size.

Combining both equations, we obtain $\delta = c^{(1/t)}$ as an empirical estimate of the discounting factor from simulated choice probabilities. This discounting factor δ is equivalent to the measure of patience defined in [58], with the caveat that they additionally privilege 'now' vs 'then' using a further one time multiplicative discount β in their model. Since β and δ do not show any difference in behavior in their subsequent analysis, we ignore this distinction in our definition.

The measure of impatience used in [59] is the implied rate of return that compensates for the inter-temporal discounting that reduces utility of later options; elicited by increasing the promissory value of the later option. In our quasi-hyperbolic discounting framework, this manipulation can be represented by $U^{(t)}(r) = \delta^t \times U^{(t)}(r) \times (1+\alpha)^t$, where α is the implied rate of return of the increased option. At certainty equivalence, $U^{(t)}(r) = U^{(0)}(r)$, implying $\delta = \frac{1}{1+\alpha}$. Hence, we can analytically obtain Dohmen's impatience measure α from our δ computation described above.

Relating working memory capacity with cognitive skill measurements The latent variable factor analysis of (Conway, 2002) establishes a correlation of 0.78 between g and performance on the Raven's Progressive Matrices (RAPM) test, and a correlation

of 0.5958 (accounting for all path coefficients) between g and working memory capacity (WMC). As (Fukuda, 2010) point out, working memory capacity can be influenced both by the simple *number* of objects a subject can hold in memory and the *resolution* with which memory instances are stored. Their factor analysis, accounting for numeric capacity and resolution capacity separately, leads to correlations of 0.83 for g vs. RAPM performance and 0.66 for g vs. the numeric aspect of working memory capacity. Since it is this latter construct that we seek to operationalize in our theory, and also since their results closely match the standard results of [70], we use the [71] study measurements to estimate a correlation between working memory capacity and performance on the RAPM. Since this correlation has never been directly measured, it is only possible to specify a range $[0.83\times0.66-\sqrt{(1-0.83^2)(1-0.66^)}, 0.83\times0.66+\sqrt{(1-0.83^2)(1-0.66^)}]$ for it. The correlation value at the low end of this range, 0.13, still predicts a positive relationship between WMC and RAPM performance, allowing us to linearly scale RAPM performance reported in the experiments of [58, 59, 57] as measures of WMC, and hence, directly comparable with our results.

Using the path coefficients from the factor analysis developed in [71], we obtain a measure of WMC as a regression coefficient weighted additive combination of the raw scores obtained on the three tests (Color, big rect k, big oval k) used in their analysis. We thereby obtain a mean value $\mu_{WMC} = 0.84\times3.36+0.88\times3.36+0.9\times3.52 = 8.9472$ and a standard deviation $(0.84^2\times0.92^2+0.88\times1.17^2+0.90^2\times1.13^2)^{1/2} = 1.64$ accurate up to scaling factors. Since the cognitive ability results in all three empirical studies [58, 59, 57] are reported in percentiles, assuming an approximately normal distribution for the ability scores allows us to map various percentile scores to notional working memory units used in our simulation. The mapping estimated and used to generate our results is displayed in Table 3.1.

WM units	Burks, 2009	Dohmen, 2010
2	1^{st} quartile	20^{th} percentile
3	2^{nd} quartile	40^{th} percentile
5	3^{rd} quartile	60^{th} percentile
7	4^{th} quartile	80^{th} percentile
9	-	100^{th} percentile

Table 3.1: Mapping of working memory size to percentile cognitive ability

Chapter 4

Cognitive efficiency as a natural action principle in decision-making

4.1 Introduction

Patterns of behavior typically considered irrational occur with such regularity that they often constitute normalcy. To illustrate, consider the choices of Mr Tiwary on a Las Vegas trip. Notwithstanding the fact that both theoretical and empirically observed odds of winning are impossibly stacked against him, Mr T patronized several different casino games. Leaving the casino, he sought out some suitable restaurant for dinner. While both nutritionally and gastronomically speaking, several inexpensive and excellent options are available, Mr T decided to splurge on an expensive steakhouse. Perusing the wine list, he pondered the multitude of options and agonized over his choice but in the end chose his favorite Merlot. While his steak was done to perfection, the enjoyment of his repast was disturbed when an animated conversation struck up between his beloved and the waiter, despite the vanishingly small probability of the conversation actually leading to any biological infidelity.

Economists have used the terms 'animal spirits' and more recently 'cognitive biases' to explain the persistence of behaviors incompatible with accepted definitions of

rationality. The observation of such 'predictably irrational' [84] behavior has reduced confidence in the conventional view of human decision-making as a rational enterprise. State-of-the-art attempts at explaining the existence of these biases typically draw upon evolutionary arguments [1, 85], tailor their explanations to explaining particular data samples and thus generate predictive, as opposed to causal, explanations. Thus, while on the hand, rational models of decision-making lie discredited through their inability to explain the existence of cognitive biases, the prominent alternative approaches towards creating heuristic-based theories are fundamentally flawed in their inability to extract generalizable causal explanations for how decision-making actually takes place. The absence of a realistic, principled theory of human decision-making is deeply problematic, since models of decision-making are central to the formulation of social and economic policies.

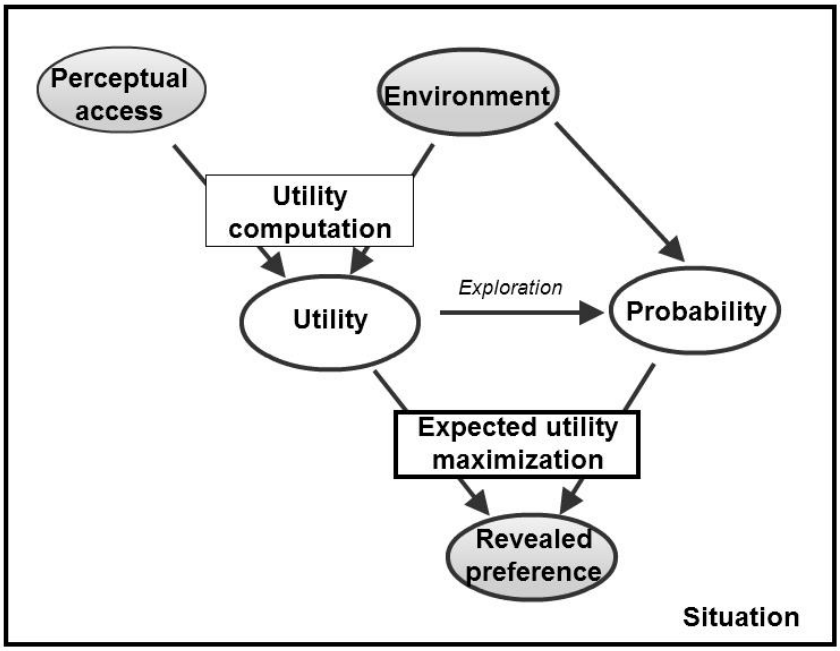
All accounts of human decision-making are formalized, explicitly or implicitly, by assumptions regarding (i) the motives and abilities of decision-makers (*agency*), and (ii) the transfer of information from the environment to the agent (see, e.g. Fig **??** adapted from [86]). Theories of decision-making must solve a number of key problems: how to represent the "goodness" of options for an agent, how an agent will select between options of varying degrees of goodness, and how the likelihood of achieving options based on the agent's actions is sensed, computed, and incorporated into choices. The pioneering work of von Neumann and Bellman provided answers to these questions that became the canonical approach, creating the framework and vocabulary for subsequent theories. Critically, the approach represents goodness via numerical utility or reward values. This representation is consistent with assuming that agent's preferences between options can be encoded as absolute numeric reward signals being embedded in the environment and that the goal of a decision-making agent is to maximize the long-term collection of this reward. These two assumptions are foundational to both *homo economicus* [87] models of economic choice and reinforcement learning [8]. As we briefly mention above, the consistent failure of such models to predict and explain real-life decisions has caused decision theorists to resort to *ad hoc* heuristics as human decision models.

While the failures of expected utility theories are incontrovertible, we believe that resorting to heuristic models, while no doubt pragmatic, cannot be the ultimate goal for

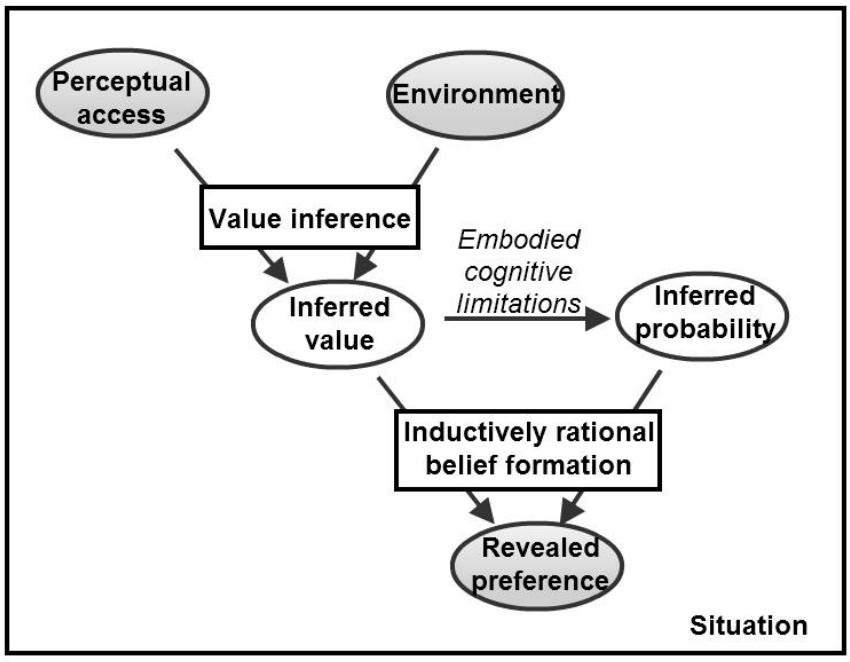
scientific theories of cognition that try to explain how humans make decisions. Therefore, in this work, we describe our effort in retrieving a rational decision theory from natural, evolutionarily motivated first principles, beginning with reasonable alternatives to the two existing assumptions described above.

In the past half-century, extensive experiments have catalogued and classified a large set of deviations from maximum expected utility decision models. Attempts to salvage the basic model have largely focus on finding generalizations of the mathematical formulations of these assumptions, leading to models that are increasingly less interpretable and generalizable. For example, the original idea of expected value was transformed to one of expected utility to account for infinite expectation paradoxes. The utility was further abstracted to ordinal utility via the von Neumann-Morgenstern (VNM) axioms [34], further abstracting the relationship between value and preference. Prospect Theory[32] proposed the replacement of standard outcome probabilities with a weighted probability function to account for differences in the way people perceive extremely low and high probabilities. These changes incrementally moved the rational choice model of decision-making away from a principled generative model to a descriptive model that can predict human choices in a limited domain of decision contexts. In response to the realization that the decisions of humans change quite dramatically depending on environmental cues governing context [1] have relaxed the requirement of a unitary basis for decision-making, suggesting, based on insights from evolutionary psychology, that humans possess a bag of different evolutionarily adapted heuristics, which they deploy in appropriate contexts. While this view has led to the creation of some well-defined and highly predictive heuristics in specialized domains, finding fundamental principles that could generate these heuristics has been elusive.

In this paper, we provide an alternative account of what constitutes rationality in human decisions. The traditional definition of rationality largely encodes the key assumptions about the fundamental motives of agents: people are completely rational when they maximize reward procurement from the environment. While reward procurement is clearly important for survival, it ignores the cost of procurement, both in terms of action, and in terms of computation. One could argue that these costs could be rolled into the utility of an option which is the target of such procurement, but that assumes that these costs depend only on the option, unrealistically ignoring the effects



(a) The canonical expected utility-based rational choice model



(b) A modified evolutionarily motivated rational choice model

Figure 4.1: This figure schematically differentiates our proposed decision theory from existing approaches. While we retain the essential rationality of selecting 'better' outcomes, we redefine rationality to reflect the evolutionary history and representational abilities of biological agents.

of context and the state of the agent on computation and action costs. Clearly, the reward value of an option also depends on context and the state of the organism because what is useful depends on the organism's current needs.

The shift away from rational analysis of decision-making has historically arisen through the realization that humans are not utility-maximizers but rather *adaptation executors*. That is, the 'biases and heuristics' research program realized [cite] that human behavior is likely adapted to environments that are substantially different from the present environment, leading to systematic deviations of choice behavior away from rational expectations. Thus, because traditional definitions of rationality were found to be deficient, it was considered more useful to try to study human behavior through enumerating adaptations that explain it. We believe, however, that such an approach is not incongruent with rational analysis, in the spirit proposed by (Anderson, 1990). Crucially, while heuristics research assumes that multiple environmental features are responsible for different adaptations, necessitating enumerative methods and positive[1] hypotheses, we believe that there is an environment that is common to all human decision-makers - their cognitive apparatus.

In this work, we show that taking elementary limitations of the human cognitive architecture into account leads to a redefinition of human rationality and a normative elicititation of preferences that exhibit both compatibility with rational expectations and stylized cognitive biases. In our revised view of rationality, the goal of a decision-making agent is to construct satisfactorily predictive theories about the relative quality of options given the organism's present need-state with as little cognitive effort as possible. The principal contribution of this paper is the mathematization of this rationality definition, yielding an alternative quantitative choice model which makes normative predictions about human behavior. As we demonstrate in our results, by extending the definition of rationality to encompass limitations of the human cognitive architecture, we retrieve a rational theory of decision-making that explains the emergence of preferences heretofore considered irrational.

[1] as opposed to normative

4.2 A cognitively efficient learning agent

In this section, we first set up the basic sequential decision problem from the perspective of an agent capable of self-reflection. Then, we determine the most evolutionarily natural objective for an agent in this setup and show that this objective, very interestingly, can also be derived from the statistical MDL principle. Finally, drawing upon these motivations, we outline the basic principles characterizing our framework and formulate them in a mathematically coherent manner.

Construe a decision to suggest the selection of a subset of outcomes out of all possible outcomes at a given time, where outcomes are cognitively separable entities in an agent's mental representation. For a series of decisions to represent a unique sequential decision task, assume the set of possible outcomes $\mathcal{S}$ does not change across decisions. Most instantiations of such a decision framework construct the agent-environment interface as utility/loss functions implicitly (or explicitly) embedded in individual outcomes [8]. However, we impute a weaker informational structure to this environment by assuming it to only possess cues that allow ecologically adapted agents to with it using internal drives and motivations to determine which outcomes are preferable.

The fundamental unit of our model of behavior is a meta-cognition-capable agent that is faced with an environment with three characteristics:

1. All phenomena in the world have dependent origination. That is, their properties and attributes, including preference judgments, are generated through the agent's interaction with them. Phenomena have no intrinsic attributes.

2. Phenomena in the environment are *transient*, which requires the agent to continually assign new preferences to outcomes

3. The agent's ability to perceive phenomena is limited

Further, our agent model is event-driven, viz. its temporality is identical with movement through its environment as a space of possible outcomes.

4.2.1 The observables

In defining an objective function for modeling realistic decisions, it is important to step away from existing paradigms of learning and decision theory, which have originated as

efforts to solve artificial problems in artificial settings. The traditional model of information transfer assumes that agents know their own preferences and that the environment provides clear signals to the reward value of option that encode these preferences. In practice, this assumption is often strengthened by positing reward signals in the environment are absolute (reward associated with one state is independent of that associated with any others), consistent (preferences do not change across time), and stationary (expected value of reward over time converges) to create a numerical representation of preferences that is invariant. These mathematical necessities are well-known to be false in practice, and several partial modeling efforts have been made to ameliorate their impact. In contrast with these incremental approaches, we question the very validity of assuming both the existence and the absoluteness of reward signals. It is unreasonable to assume the information transfer between options and their values is so transparent. How much utility that an option affords towards satisfying a need is seldom known precisely by an organism, and then only with considerable experience. We argue that utility of an option is better construed as a prediction that an option will better satisfy an organism's needs relative to the other options available. In summary, rationality based on maximizing reward procurement fails to account for the costs, limitations and needs of actual organisms.

Where does reward come from? We claim that the ability of an agent to infer reward arises from the relative goodness of various outcomes to the agent across its evolutionary history. The actual reward inferred, however, arises from the agent's personal history of interaction with outcomes[2] . Further, we suggest that rewards cannot be stores of absolute value, but can only be understood as comparisons between various outcomes. Assuming the absence of counterfactual reasoning ability, our model of information transfer from the environment to the agent assumes rewards to live in an affine space and to emerge upon the activation of a particular set of outcomes by the agent during its exploration of the environment.

Also, while most sequential learning tasks track long-run reward/loss, we consider tracking immediate performance as a more natural setting for establishing existential

[2] For example, a human may have a predilection towards liking fatty food because of the survival value of fatty food to the genotype. However, the actual value of various fatty foods to this particular human will depend on her personal experience with them.

goals. We expect a successfully evolutionarily adapted meta-cognitive organism to possess an intelligence directed towards satisfying two goals (i) orienting beliefs accurately with respect to the environment and (ii) increasing confidence in the quality of beliefs.

Now, in comparing the two goals defined above, an interesting observation emerges: while both are equally important in principle, it would appear that satisfying the second 'interior' goal, with exterior referents not necessarily needed for validation, could prove to be easier than satisfying the first one, which would depend entirely on the predictability of the agent's environment. In an evolutionary situation where two goals are necessary, one is easier to optimize than the other, and the sum total counts towards the survival fitness function, a natural hypothesis would be that organisms finding optimal solutions to the interior goal while maintaining satisficing solutions for the exterior goal would be selected for.

From these qualitative deductions, we hypothesize that *biologically realistic predictive decision-making is fundamentally a meta-cognitive heuristic of optimizing a self-perception of improved predictive ability.* A mathematical model of decision-making can, therefore, be expressed as optimizing some combination of both the agent's interior and exterior goals,

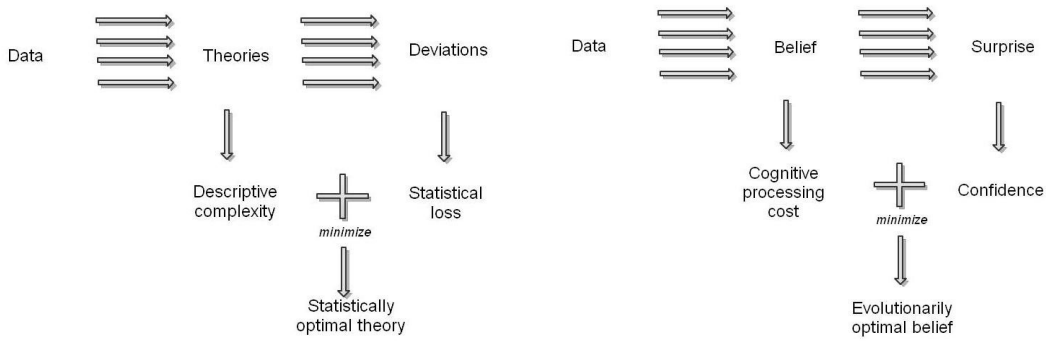
$$\underset{\mathbf{x}}{\operatorname{argmin}}\ T + C^{-1}, \tag{4.1}$$

where, hypothetically, T quantifies the meta-cognitive interior goal-object representing self-perceived 'cost' interpreted as an inverse of 'ability', C quantifies the exterior goal of quality of predictions, and x quantifies the agent's revealed belief about the relative quality of experienced outcomes and thus constitutes a stochastic decision.

In a striking parallel with the success of structural information-theoretic approaches in modeling visual perception [88], posing the meta-cognitive intelligence framework as one of optimal learning brings the goal of our biologically inspired metacognitive agent extremely close to the goals of information-theoretically optimal learning agents. Specifically, if optimizing self-perception of improved predictive ability can be equated with confidence in the ability to construct better theories, our hypothesis is almost identical with claiming a minimum description length (MDL) [89] basis for biologically realistic decision-making.

In general, an MDL agent has two goals, to construct a compact theory and to minimize its prediction errors. For instance, consider a data set that currently requires

Z bits to describe completely. Now assume that there are i possible theories explaining this data, each theory itself requiring X_i bits to describe and the deviations of each theory's predictions from the real data requiring Y_i bits to describe. MDL proposes[3] that the best theory is the one that minimizes $X_i + Y_i$. The basic intuition in MDL is that the explanatory value of a theory is a consequence of its data compression ability, i.e. a good theory will achieve a low $\frac{X_i+Y_i}{Z}$ ratio.



(a) A schematic of obtaining a good theory from data following a minimum description length approach.

(b) A schematic of obtaining a good belief from environmental stimuli using an evolutionary MDL principle

Figure 4.2: This graphic outlines the homologous nature of an evolutionarily optimal meta-cognitive decision strategy with minimum description length principles.

These goals are similar to the ones we have proposed for meta-cognitive intelligence. In our case, the goal of the agent is to construct a belief explaining sequential inputs of environmental stimuli. Our cognitively efficient agent can be seen to differ from an MDL agent in two significant ways, (i) it will operate in an online setting and so must have a different notion of compactness than the complexity measures used in MDL and standard machine learning and (ii) instead of minimizing prediction errors, it will try to bring them down to a satisficingly [90] low level. Let us now flesh out the correspondences and differences in some greater detail.

We assume that our agents can obtain beliefs with respect to environmental stimuli. With respect to the decision problem, we consider a special class of beliefs about the

[3] With several philosophical qualifiers; see [89] for a detailed discussion.

quality of outcomes to be salient to our construction. We call these beliefs 'quality-beliefs' to differentiate them from the more general concept. Quality-beliefs are represented in our framework as probability distributions over the space of possible outcomes and are taken to reflect the agent's action preference stochastically. Given a particular quality belief $\mathbf{x}_t$ at time t, an agent will prefer a set of actions it believes lead to the superior outcome with a probability $p(y/\mathbf{x})$. The quality-belief, in effect, parameterizes the agent's interaction with its own internal states.

It is quite easy to see that 'quality-beliefs' are the cognitive equivalents of statistical 'theories' from the stand-point of a learning algorithm. They are learned from environmental data and their accuracy reflects the agent's ability to accurately predict and embark upon actions best oriented with its current goals. However, as with statistical theories, the agent's 'quality-beliefs' will fail to predict the optimal course of action in the future with complete accuracy. Upon facing instances of deviation from its predictions, the agent will experience some degree of 'surprise'. The surprise experienced by an agent operating with quality belief $\mathbf{x}_a$ with respect to another quality-belief $\mathbf{x}_b$ can be quantified with an information divergence [69] of the form,

$$R(\mathbf{x}_a, \mathbf{x}_b) = \sum_{j=1}^{n_a} \mathbf{x}_a^j(s) \log \frac{\mathbf{x}_a^j(s)}{\mathbf{x}_b^j(s)}. \tag{4.2}$$

The information divergence measure is intuitively suitable for representing differences between beliefs encoded as probabilities, since it is asymmetric and non-metric. The asymmetry leads to the current belief being privileged in a particular way (easy to find past belief that is closest to current belief, converse is hard). The non-metric nature of the information divergence allows for intransitive selection between gambles, as seen in the Ellsberg paradox, for instance, to occur.

4.2.2 Cognitive processing cost

As we have hinted earlier, our model uses a different measure for evaluating the 'goodness' of a theory than classical MDL. Whereas the latter approach, and, in fact, most statistical regularization techniques use information-theoretic complexity measures on a theory as a proxy for the uncomputable Kolmogorov complexity, such an approach is judged unsuitable for our purpose for two reasons. First, the agent acquires data

sequentially and, in general, has no long term goal and hence, reason to combine data from multiple observations. Each data instance, and the quality-belief generated therefrom must be evaluable in its own right. Second, while descriptive complexity is an elegant measure for a theory's generalizability and power in a statistical setting, it does not have a cognitive motivation and need not be suitable for an evolutionarily motivated learning agent. Thus, as our first significant departure from the statistical norm, we replace the notion of descriptive complexity with one of cognitive processing cost as a measure of the goodness of a particular quality-belief.

Our conception of processing cost resembles Conlisk's notion of 'deliberation cost' [91]. The difference lies in the fact that while Conlisk [91] and Russell [92] consider calculative or deliberation cost as a cost of utility computation, our definition assigns it a different and more neuronally justifiable intepretation. We posit that a realistic agent would possess a finite 'bandwidth' of this resource. Also, the number of decision instances requiring the use of this resource would be unknown and, depending on the rate of experience, could subjectively appear to be very large or very small. The agent, therefore, would be best served in trying to improve or maintain its efficiency in deploying cognitive processing for decision-making. Thus, it would expect to minimize its cognitive processing cost.

But how does one describe processing cost in a quantifiably meaningful manner? While, in general, processing costs will depend on the agent's computational architecture and will vary based on the agent's particular processing algorithm, we focus here on developing a generic plausible notion of processing costs for biological and neurally inspired agents.

For an agent to determine a new *quality-belief* for making choices in its environment it needs to (i) process information about the quality of available options at the present time and (ii) recall quality-beliefs generated when handling decision instances considered salient to the present one. Now, processing costs corresponding to the former computation do not fall within the agent's control, as they are embedded in the agent's perceptual apparatus. Thus, the only variable processing cost corresponds to the agent's memory recall operations. Naturally, the variability of cost will depend on the number of prior quality-beliefs recalled for use in constructing the new quality-belief, making this number the variable of interest in defining processing cost.

With this argument, we have reduced the general concept of computing cognitive processing costs to one of computing the cost of an agent's memory recall. This cost, of course, would also depend on the memory access model that we choose to construct. Following our MDL intuition, we hypothesize that evolutionarily optimized agents are likely to assign storage space/access to past quality-beliefs in a way that maximizes information compression. Recalling coding theory [93], in order to construct optimal encodings, frequently occurring symbols are assigned shorter codes and infrequently encountered symbols are assigned longer codes. We postulate a similar mechanism at play for memory access, arising evolutionarily in order to promote efficient memory use. Thus, we hypothesize that quality-beliefs corresponding to informationally exceptional decision instances will take up disproportionately larger storage space in an agent's memory. Assuming *a priori* spatially random access via a sampling without replacement procedure, such quality-beliefs will be easier to retrieve from all possible candidates at a particular retrieval instance[4] .

Given this particular model of memory access, a natural quantification of cognitive costs follows. Let $s \in \mathcal{S}$ represent possible outcomes in the decision task. The statistical departure from the norm that would define an exceptional quality-belief is quantified using the measure of surprise defined above in 6.2. The informational exceptionality of a past quality-belief $\mathbf{x}_{\text{old}}$, and hence the ease with which it will be available for recall to the agent, can then be quantified as,

$$A(\mathbf{x}_{\text{old}}) = |R(\mathbf{x}, \mathbf{x}_{\text{old}}) - \overline{R}|, \tag{4.3}$$

where $\overline{R}$ is the average of surprise experienced in the past by the agent and x is the agent's current quality-belief. This quantity will be high for values of surprise that deviate from the average value to either extreme.

Given this measure of ease of memory access for each past quality-belief, a reasonable measure of the processing cost of selecting a subset $\mathcal{M}'$ out of the set $\mathcal{M}$ of all past quality-beliefs would be the inverse informational exceptionality-weighted sum of the nominal cost of accessing all quality-beliefs in $\mathcal{M}'$. Assuming the nominal cost of

[4] To understand this intuitively, consider a stick of length 10 units, with 7 units colored red, 2 colored green and 1 colored blue. Locating a red unit via random unit selection on this stick will naturally be much faster than a blue unit.

accessing each quality-belief to be unity, the total cost of memory access T would be,

$$T = \sum_{\mathbf{x}_i \in \mathcal{M}'} A^{-1}(\mathbf{x}_i), \tag{4.4}$$

While it is necessary to grant us our particular model of memory access in order for this derivation to make sense, the eventual model of memory access cost appears to be quite natural. In effect, exceptional past quality-beliefs are easier to access than unsurprising ones. Note, however, that the manner in which we have defined informational exceptionality means that both quality-beliefs that led to a high degree of surprise as well as ones that led to a very low degree of surprise are considered exceptional. This corresponds to the natural intuition that being extremely right is just as exceptional for a cognitive agent as being extremely wrong.

As we have discussed above, the cost of memory access is the only portion of the total processing cost under the agent's cognitive control. Thus, in the context of an optimization problem, it is identical with the agent's total cognitive processing cost for a given decision instance. Given sequential access to data and absent long-term goals, we consider this quantity to be a realistic measure of a theory's 'goodness' from an information-theoretic standpoint. Interestingly, we are not the first to suggest an alternative measure of 'goodness' in online settings. [94] has proposed using a speed prior ensuring that computations that end earlier are more preferable in a complexity-theoretic setting which shares our underlying motivations, but not the biological computational substrate.

4.2.3 Defining confidence

Our final correspondence with the statistical MDL framework arises from an evaluation of the measure of deviation from predictions in both settings. While learning theoretic systems use measures of loss and/or expressed as the difference between the predicted value and the observed 'true' value, our task is complicated by the fact that no objective truth values can be posited to exist in a realistic decision-making environment.

Rather than assume the existence of objective true values embedded in the environment, we define a somewhat more sophisticated concept of *reward-inference* as an interface between the agent and the external environment. At a given decision instance,

an agent observes possible outcomes through the performance of particular actions from its current state. It assigns some preference to actions it *perceives* as being more desirable than others. Not all possible outcomes need be observed, and only outcomes that are observed are evaluated. For a set of possible outcomes $\mathcal{S}$, we encode the agent's perceived preference with respect to environmental outcomes as a probability distribution $g(s)$ over the outcome space. Since obtaining this distribution, in general, requires some normative processing of environmental stimuli on the agent's part, we call this quantity *reward-inference*, identical with the interpretation we assigned to relative desirability in Chapter 2. In this chapter, however, we do not enter into the details of the inference process, assuming only that an agent perceptually adapted with respect to its environment is able to consistently prefer the better out of experienced outcomes. Maximizing perceived reward is thus, equivalent to accurately predicting future reward-inference. This construction allows us to model external motivations for an agent's actions without having to postulate oracular access to optimal behavior, as is the case in almost all existing decision learning approaches.

Intuitively, it should be clear that a 'good' quality-belief would correspond to a scenario where the quality-belief corresponds accurately with future reward-inference $g(s)$. We now define a quantitative measure of performance with respect to the 'goodness' of quality-beliefs. For an agent that is not maladaptive, improved reward perception will arise when (a) the agent will have constructed a history of successful prediction and (b) the agent's current quality-belief will be relatively unambiguous. Since we already possess the quantity *surprise* as an inverse indicator of predictive success[5] , we can conceive of a measure of the agent's confidence in its ability to gain reward as an inverse function of both uncertainty experienced with respect to the current quality-belief and cumulative surprise experienced with respect to past quality-beliefs. Basically, when an agent finds itself in possession of a stable and unambiguous quality-belief, we expect its *confidence* to increase. We therefore define a cognitively efficient agent's predictive *confidence* $C : \mathbf{x} \to [0, 1]$ as,

$$C = \frac{1}{C_{\max}} \frac{\log |\mathbf{x}| - H(\mathbf{x})}{\sum_{\text{memory}} R(\mathbf{x}, \mathbf{x}_{\text{old}})}, \tag{4.5}$$

[5] Note the difference between predictive and ecological success. For example, someone saying "What a pleasant surprise!" is referencing his/her inability to predict an ecologically useful outcome.

where the numerator, as may be readily observed, is anti-monotone with respect to the Shannon entropy of the quality-belief. Note that C is to be normalized with respect to the greatest value it has previously been observed to achieve.

The connection between an agent's confidence defined above and its anticipation of the goodness of its beliefs is not entirely evident at first sight. Consider a scenario wherein, during a preliminary learning period, in a majority of cases, the reward-inference indicates the outcome s_i as overwhelmingly more preferable. This suggests that the agent's quality-beliefs, being aggregations of reward-inference samples, will also reflect this preference towards s_i. Now, while the agent is exploring its environment, it may encounter a similar or different reward dynamic. If it encounters the same environment, its existing quality-belief will be well-adjusted with respect to the incoming reward-inference, and the information divergence between the updated quality-belief and past quality-beliefs will remain low. Thus, in a familiar environment, the agent's confidence will increase as it receives a steady flow of perceived reward. In a different environment, the agent's preference for s_i may prove to be maladaptive; the new reward-inference will force its quality-belief response to differ from its past quality-belief, creating instances of greater surprise. The transition to a new environment, characterized by a period of maladaptation and diminished reward access followed by gradual adaptation, leads to a drop in confidence. Confidence is thus a measure of the agent's anticipation of the goodness of its future decisions. Trying to maximize goodness of future decisions is therefore identical to trying to maximize confidence.

4.2.4 The objective

We now possess all the conceptual entitites necessary to define an objective for a realistic meta-cognitive decision learning agent following MDL principles. An objective function of the form (4.1) would appear to be indicated. However, in our second departure from the traditional MDL formulation, we assume, for reasons detailed in Section 4.2.1 that it is more realistic for agents to attempt to minimize cognitive processing costs (the interior goal) while maintaining a satisfactory level of predictive ability (the exterior goal) and

hence, confidence. The cognitive efficiency objective function, therefore, takes the form,

$$\begin{aligned} \underset{\mathbf{x}}{\text{argmin}} \quad & T \\ C_{\text{new}} \quad & \geq C_{\text{old}}. \end{aligned} \tag{4.6}$$

4.3 A natural solution

Having set up our decision problem in the form of a constrained optimization problem in (6.2), we now turn to the less well-defined issue of developing an algorithmic solution to this problem that is both (i) provably optimal and (ii) biologically plausible. While the first of these conditions is easy to verify mathematically, the second is harder to rigorously validate and depends necessarily on the reader's judgment.

At each decision event, the agent has access to two inferential processes - one at the cognitive level allocating processing resources, which we model as an internal memory update, the other at the conceptual level, obtaining reward-inference signals from the environment in a statistically optimal manner. Hence, for the purpose of our algorithm, we consider the agent's quality-belief to have two sources of information: the discrete probability distribution $g(s)$ represents the reward-inference signal obtained perceptually from the environment, which may or may not be present at each and every decision instance. The memory $m(s)$, likewise a discrete probability distribution, is determined from prior quality-beliefs and is updated via a mechanism that we describe below.

4.3.1 The cognitive algorithm

Since the only variable cognitive processing costs are associated with memory, our cognitive algorithm mimics recall operations in the agent's memory, where this memory is constructed such that events with high informational exceptionality are more easily accessible than those with low informational exceptionality and thus are less 'costly' to access. Availability is defined as shown in (6.3) with the caveat that the *average* surprise $\overline{R}$ is now computed not over all past decision instances, but over a privileged subset of them. This privileged subset, which we call the 'salient set' $\mathcal{K}$ corresponds intuitively with the active memory of the agent, is composed of a sufficiently large number of highly available past decision instances and is generated by a resampling of prior

decision instances using informational exceptionality as the selection criterion.

Algorithm 1 Algorithm for salient set construction

```
Input ← 𝒢.
x_0 = m_0 = U(|g_0|).
0 ← C_0, C_1.
0 ← R̄.
{∅} ← 𝒦.
for i = 1 to |𝒢| do
  for j = 1 to i do
    Compute R(x_i, x_j) using (6.2).
    Compute A(x_j) with respect to x_i using (6.3).
    if ∃x_k ∈ 𝒦, A(x_j) > A(x_k) then
      x_j ∈ 𝒦.
      Compute m_i using (4.7) and x_i using (4.9) with m_i and g_i ∈ 𝒢.
      Compute C_i using (6.5).
      if C_i > C_{i−1} then
        x_j ∈ 𝒦.
      else
        x_j ∉ 𝒦.
      end if
    end if
  end for
end for
```

In other words, the memory constitutes the agent's recollection of quality-beliefs it has used in the past for making decisions it considers to be salient to the present decision. Memory construction in our framework is modeled as an average over all past quality-beliefs considered salient to the current decision,

$$m_i(s) = \frac{1}{\max(1, |\mathcal{K}|)} \sum_{k=1}^{|\mathcal{K}|} \mathbf{P}_k \; x_k(s) + \mathbf{N}_k \; \overline{x_k}(s), \tag{4.7}$$

where $\mathbf{P}$ is an indicator vector that takes values 1 for low regret salient instances (zero otherwise) and $\mathbf{N}$ takes value 1 for high regret salient instances (zero otherwise). The notation $\overline{m}$ represents inverting the set of preferences under consideration and can be obtained in several ways, e.g. subtracting each component value from 1 and renormalizing. In cases where the salient set is empty, m_i simply takes on the value of the immediately prior quality-belief x_{i-1}, reflecting the intuition that no memory recall took

place.

4.3.2 Combining memory and the environment

If we assume memory and its cognitive mechanisms to be a black box, the remaining sequential decision task for an agent becomes identical to online density estimation problem studied in the machine learning literature. In online density estimation, an algorithm receives a sequence of data vectors. The algorithm uses its current parameter settings to predict the value of the next data vector. After making the prediction, the algorithm receives the next data point and incurs a loss determined by a loss function dependent on the algorithm's prediction and the actual data value. The algorithm uses this loss to update its parameter settings. Now, observe that the prediction task accomplished in our reward-inference process is identical with that faced by an online density estimator - finding an ecologically suitable action policy given sequential environmental inputs. In this case, the agent's quality-belief takes the place of a parameter setting, and is updated given a new reward-inference signal. The stochastic link between quality-belief and the agent's action, which we have earlier introduced as the density $p(y/\mathbf{x})$ can be ascribed any parametric form. Without significant loss of generality, we consider it to belong to any exponential family distribution [95]. Given this setup, Azoury & Warmuth [96] have shown that the MDL-optimal update of the expectation parameter, in terms of minimizing relative loss compared to a batch algorithm given access to all data in advance, can be written as

$$x_{t+1} = \eta_t \eta_{t-1}^{-1} x_t + \eta_t g_{t+1}, \tag{4.8}$$

where, η is an algorithm specific learning parameter and $\eta_t \eta_{t-1}^{-1} + \eta_t = 1$. In other words, the optimal quality-belief update is a convex sum of the existing quality-belief with the new reward-inference signal, parameterized via a learning rate.

From the analysis above, we obtain the intuition that a statistically optimal mechanism for updating our agent's quality-belief would involve a convex sum between the existing quality-belief and the incoming reward-inference signal. Since the existing quality-belief at every decision instance is identical with memory, at the i^{th} decision instance, the current agent quality-belief can be calculated as

$$x_i(s) = C_i m_{i-1}(s) + (1 - C_i) g_i(s), \tag{4.9}$$

thereby satisfying the constraint imposed by (4.8). Note that we have endowed the learning rate η with a predictive confidence interpretation. This follows the intuition that an agent with high confidence in its predictive ability will trust its own judgment more than external input. On the other hand, an agent with low confidence with trust external inputs more than its existing beliefs.

(4.7) and (4.9) together constitute the overall quality-belief update equation for our algorithm. The cognitive processing optimality criterion is introduced and solved through the problem of optimal salient set construction. The quality-belief 'goodness' criterion is introduced through the confidence constraint imposed on the salient set construction procedure, viz. the set is updated only if doing so improves the agent's predictive ability with respect to the environment.

Thus, our biologically-inspired approach to solving the original cognitively efficient decision problem is to minimize T with respect to salient set membership if doing so leads to an increase in the agent's predictive confidence,

$$\begin{aligned} \underset{\mathcal{K}}{\operatorname{argmin}} \quad & T \\ & C_i \geq C_{i-1}. \end{aligned} \tag{4.10}$$

It is interesting to note that the requirement to maintain confidence makes our model asymmetric with respect to low and high regret instances by skewing the agent's preferences in favor of low regret salient sets. This corresponds to the intuition that an agent should much rather prefer being right to being wrong and, amongst equally high informationally exceptional instances, would preferentially select low regret quality-belief instances over high regret quality-belief instances to populate its salient set.

Assume that $\mathcal{G}$ is the set of all instances of environmental feedback that have been or will be encountered by our agent in a series of decisions[6] . Given $\mathcal{G}$, algorithm 1 presents a straightforward way for finding optimal $\mathcal{K}$ while respecting the confidence constraint. The set $\mathcal{K}$ is then used to construct the memory update 4.7. The memory update, along with reward-inference and confidence is used to construct the optimal quality-belief for the relevant decision instance using (4.9).

We freely confess our inability to meaningfully assess the validity of our algorithmic approach in light of recent developments in understanding of neuro-biological processes

[6] Naturally, this set will be populated incrementally in practice.

involved in decision-making, and would be very happy to incorporate both positive and negative empirical support (and implications thereof) for the existence of a neurophysiological substrate for our algorithm in future iterations of this work.

4.4 Explaining cognitive biases

While simply constructing a choice-learning algorithm capable of quantifying internal motivations would pass for an interesting theoretical exercise, in this section, we present evidence to show that this construction naturally leads to realistic quantitative prediction of putatively *irrational* behaviors empirically observed in human subjects.

While a very large number of such biases have been observed and documented, in this study, we concentrate on three 'families' of biases that have been shown to (i) exist independent of framing and context and (ii) subsume a number of other cognitive biases.

Specifically, we replicate experimental results from three classic studies - Kahneman and Tversky's demonstration of probabilistic sub-additivity as a violation of the independence axiom of expected utility theory in human subjects [32], Klayman and Ha's explication [97] of the nature of Wason's experiments on confirmation bias [98] and Deese and Kaufman's demonstration of serial position effects in memory recall tasks [99].

4.4.1 Risk aversion

Kahneman and Tversky [32] proposed prospect theory largely to explain deviations from expected value predictions in certainty-equivalence studies on evaluations of risky prospects in human subjects. They observed that subjects consistently exhibited a four-fold pattern of behavior when confronted with risk: risk-seeking for gains with low probability, risk-aversion for gains with high probability, risk-seeking for losses with high probability, risk-aversion for losses with low probability. [48] explain the emergence of this pattern as a consequence of the disproportionate weighting of low-probability outcomes in human subjects[7] .

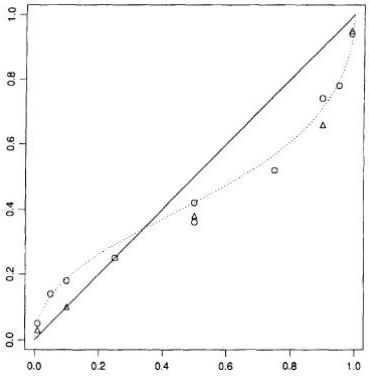
[7] This explanation was subsequently amended in [32] to restrict over-weighting only to 'extreme' low-probability events as opposed to all low probability events. Note that this development is naturally accounted for in our model.

The experimental setup for their experiments is fairly straightforward: subjects are asked to select between a 'safe 'gain/loss prospect of known value and one of unknown value determined as a Bernoulli choice between two known outcomes. For example, a subject could be asked to choose between selecting a prospect that pays \$0 with a probability of 0.9 and \$50 with a probability of 0.1 and a set of prospects guaranteed to pay anywhere between \$2 and \$20 (say). The subjects were required to indicate their preference between the risky and safe prospects for all the safe prospects presented to them. The certainty equivalent value was estimated as the midpoint between the lowest accepted and the highest rejected value from among the safe prospects. Selections where the certainty equivalent value exceeded the expected value of the risky prospect (\$5 in this case) were considered risk-seeking, while those that were lower were counted as risk averse.

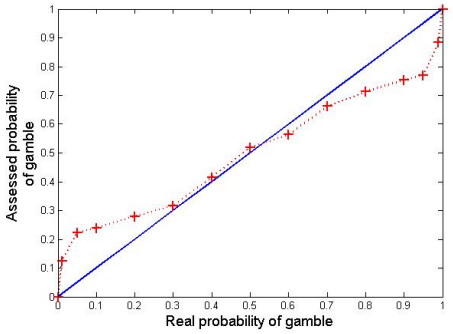
In order to simulate the experimental setup described in [32], we design our outcome space to consist of two possible outcomes: select safe prospect or select risky prospect. For every decision instance, the payoff for the risky prospect is sampled from a Bernoulli distribution appropriate for the gamble. For the gamble in the example above, this means that the risky prospect will pay \$0 in about 9 out of every 10 decision instances. The reward-inference signal is constructed to assign a preference of 1 to the better prospect (and 0 to the worse prospect) at every instantiation. Thus, a choice between a gamble with a 0.1 probability of paying off against a certain safe outcome is modeled as a generative mechanism for reward-inference that reflects a selection [0 1] biased towards the safe choice 90% of the time and the alternate risky choice [1 0] 10% of the time.

We provided each one of a population of 200 agents with a series of 100 such reward-inference signals. A series is presumed to indicate the 'learning' phase for an agent with respect to a particular choice problem involving risk evaluation. At the end of a series, the agent is assumed to possess, in the form of its final preference, an evaluative model for selecting between the prospects offered in the [32] selection task. We modify the probability of winning or losing the gamble by modifying the Bernoulli distribution parameterizing the reward inference distribution.

In Fig 4.3, we see that our simulation replicates results that are qualitatively similar to the experimental results obtained from human subjects in [32]. Remarkably, agents



(a) Results from experiments on human subjects attempting to find subjects' implicit certainty-equivalence with respect to gains/losses and its deviation from mathematical expectation. Historically, this was the predominating motivation for the development of prospect theory.



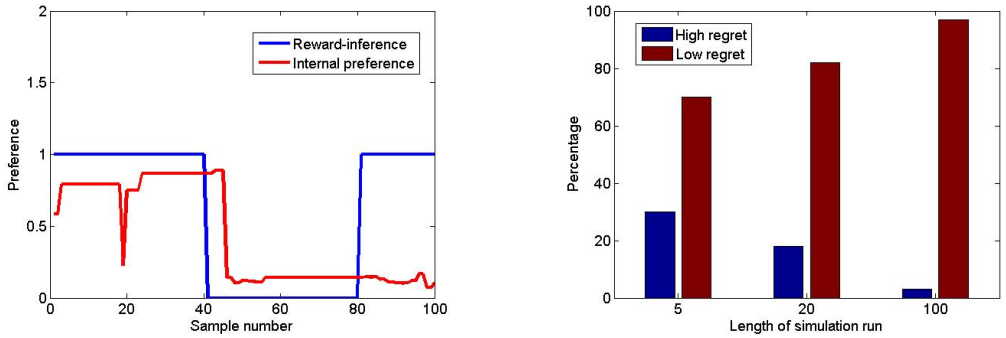
(b) Results from simulation of prospect theory experiment using cognitively efficient agents as subjects. The blue line represents the idealized expected value prediction while the red markers indicate average preference of 200 agents having experienced a history of repeated exposure to a choice selection task between a risky gamble with a certain (x-axis) probability of succeeding and a safe choice.

Figure 4.3: Cognitively efficient learning generatively reproduces experimental results described via prospect theory

running our cognitively efficient learning algorithm consistently present the same four-fold pattern of risk aversion observed in human subjects. This leads us to hypothesize that the biases documented by Kahneman and Tversky, which have subsequently motivated the development of prospect theory and other generalized expected utility theories are, in fact, adaptive in nature rather than existing *a priori* in human decision-makers. Our model presents, to the best of our knowledge, the first generative mechanism for estimating and potentially quantifying Kahneman and Tversky's four-fold pattern of risk aversion.

4.4.2 Confirmation bias

The term 'confirmation bias' often references biased hypothesis evaluation, differential memory recall, belief divergence, attitude polarization and other biases arising in different experimental contexts. The fundamental similarity shared by all these biases is the tendency for subjects to prefer information that confirms their existing preconceptions/hypotheses over objective evidence.



(a) Cognitively efficient agent evidencing confirmation bias

(b) Percentage of high and low surprise instances active in agent's salient set for runs of different lengths.

Figure 4.4: Different flavors of confirmation bias exhibited by cognitively efficient agents

Fig 4.4(a) displays typical performance of a cognitively efficient agent on a binary prediction task. Given consistent reward-inference favoring one outcome (say $\{0, 1\}$), the agent's preference for this outcome increases, which is entirely rational. Then, consistent reward-inference favoring the other outcome $\{1, 0\}$ is provided, causing the

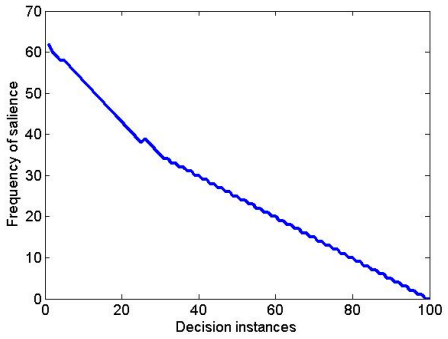
agent to reverse its preference (after a brief delay), which, again is entirely rational. However, when the reward-inference is switched yet again back to the original outcome, the agent does not switch, but continues to confirm its recent preference for the other outcome (see, e.g. Fig 4.4(a)). This superficially irrational behavior follows naturally from the tendency of our agent to retain its existing theory if formulating a newer theory would cause its predictive confidence to drop.

The first scientific evaluation of confirmation bias is historically assigned to Wason's [98] rule-discovery experiments. However, Klayman and Ha [97] proved that what Wason had actually shown was that human subjects prefer using positive test strategies, i.e. instead of trying to find counter-examples to a hypothesis, they seek to validate it. Interpreting these findings in our framework, observe that a falsificatory negative test strategy of trying to rigorously disprove a held hypothesis would create several high surprise decision instances for a cognitively efficient agent. Conversely, deploying positive test strategies would create (given a predictable environment) low surprise instances. Since part of the agent's goal is to maximize its expectation of future reward, and since this expectation, in the form of confidence, will vary inversely with the cumulative surprise in the agent's recalled history, it will strongly prefer making choices that lead to low surprise, and hence will prospectively prefer positive test strategies. Fig 4.4(b) shows the agent's preference for low surprise decision instances. Very interestingly, we find that the agent's preference for positive test strategies appears to emerge gradually as it becomes more sure of its existing hypothesis. This corroborates the information-theoretic intuition [97] that such a preference arises as an information-processing response to environments where positive queries have higher informational content than negative queries.

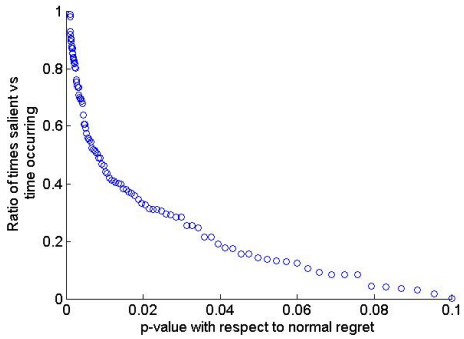
4.4.3 Ordering effects

When asked to recall a list of items in any order (free recall), subjects tend to begin recall with the end of the list, recalling those items best (the recency effect). Among earlier list items, the first few items are recalled more frequently than the middle items (the primacy effect). While experiments on human subjects have primarily used mnemonic memory recall as their domain, it is suggested that the recall process in the verbal domain is not likely to be very different from that in the domain of past beliefs, which are retrieved in

our case using the cognitively efficient learning algorithm. However, word recall does not involve choice selection after each input stimulus, so, to remove this extra dimensionality from the cognitively efficient learning problem, we let agents track a constant reward-inference signal. To measure the likelihood of recollection, we simply count the number of times different past quality-beliefs are selected as salient for any decisions within the run. Assuming that the frequency of recall is proportional to ease of recollection in the absence of predictive cues, we see in Fig 4.5(a) that our cognitively efficient learning algorithm presents a very pronounced primacy effect. Outcome extremity is measured here as a p-value with respect to a normal distribution of surprise with mean and variance determined by empirical values obtained in the current run.The ratio of membership against occurrence is simply the number of times a decision instance assigned the statistical rarity described above is selected for salient set membership to the number of times such instances actually occur during a learning run.



(a) Number of times a cognitively efficient agent tracking a constant reward-inference signal considers the x^{th} quality-belief encountered as suitable for constructing a new one. Results averaged over 100 different runs with 100 decision instances each.



(b) Scatter plot of outcome extremity (measured as a p-value) vs ratio of a quality-belief's membership in salient set against total instances of its occurrence across 100 trials of 100 decisions each.

Figure 4.5: Ordering and 'peak-end' effects in cognitively efficient learning.

Algorithmically speaking, this occurs because the surprise statistic used to determine salient set membership is extremely volatile at the beginning of every run and gradually settles down once the agent has acquired an optimal quality-belief with respect to incoming reward-inference. The early volatility causes several surprise instances to appear sufficiently different from past instances to classify them as salient. By the later stages

of a run, assuming the reward-inference signal does not change, the reward statistic is stable and fewer instances sufficiently far from the mean are detected and classified as salient.

Note that our algorithm does not appear to show a recency effect [100]. We observe that this is because we have not sought to model the dynamics of belief retrieval. To take a simple example, if an agent is posed a decision problem *after* it has been presented with a set of environmental cues, it is more likely to recall them in an inverse order of presentation to the one we have used in our experiments. Such a *retroactive* agent would then show recency bias as opposed to primacy bias. Therefore, a more pointed criticism of our model would be that it fails to reproduce both primacy and recency biases with the same environmental structure. It should be pointed out, in this context, that the existing literature on serial ordering effects is divided [101] over the possibility of replicating both primacy and recency biases within the same decision context. Our findings, therefore, do not contradict existing results. Further modeling the order of belief retrieval and its effects on recall frequency presents an interesting future direction from this work.

Finally, we note that our algorithm empirically supports the observation [102] that the value of past experiences is assigned through evaluating them at their peaks, not on an average. In our case, the peaks are determined, not in terms of absolute reward, but in terms of surprise experienced by the agent. Instances corresponding to extremely low or extremely high surprise predominantly influence a cognitively efficient agent's internal preferences with respect to a sequence of events (see Fig 5.2(b)). A retroactive cognitively efficient agent, therefore, will display behavior analogous to peak-end effects.

4.4.4 The Technion prediction competition

While demonstrating biases in an abstract sense is theoretically useful, we also subject our algorithm to a more practical test, by attempting to replicate human behavior on a variety of certainty equivalence choice tasks obtained as part of the Technion prediction tournament organized by Ido Erev and Alvin Roth in 2008 [15].

Data description: The Erev-Roth prediction competition collected subjects' preference for risky prospects vs a safe payoff for a selection of 60 prospects, covering a broad range of risk probabilities and both gain and loss paradigms. Data was collected

under three different experimental settings: one-shot choices with described prospect probabilities, repeated choice, and single choice after sampling. We briefly describe the experimental settings for each of the three conditions below:

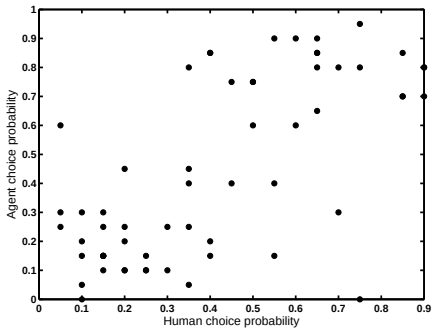
- **Description:** In this condition, 20 subjects are asked to select between a risky option and a safe option for 60 different problems, where the probability of the risky option paying off is deterministically given to the subject before they choose.
- **Experience with repetition:** In this condition, 100 subjects are divided evenly into 5 cohorts, each of which is presented with 100 trials of 12 different problems, with the risky option's payoff probability unknown. Subjects choose between the risky and the safe option, and receive feedback in the form of payoff for each trial.
- **Experience with sampling:** In this condition, 40 subjects are divided into two cohorts, each of which is presented with 30 different problems, with the risky option's payoff probability unknown. For each problem, subjects are allowed to indefinitely sample trials from both the risky and the safe options, receiving payoff information for each sampling trial. After having sampled sufficiently, subjects are asked to choose between the risky and the safe option with finality. Making a choice concludes the presentation of that particular problem.

Of these, only the repeated choice paradigm maps directly to the sequential learning setup of our algorithm. Therefore, to obtain predictions from cognitively efficient agents in each of the three conditions, we interpreted the Erev-Roth in condition-specific ways to obtain choice-event sequences in each of the three cases.

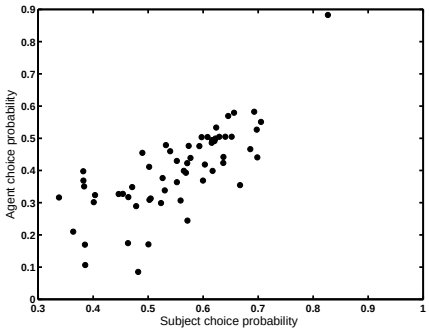
- **Description:** We assumed this condition to be equivalent to prior exposure to a history of N=200 instances of the same gamble, with an observed win-loss frequency that matches the described win probability of each gamble
- **Experience (repetition)**: This regime remains unmodified, since it directly gives us a choice event sequence.
- **Experience (sampling):** Here, we assumed that each sampling instance is a choice trial, with the actual choice instance simply the final choice in each trial sequence.

Obtaining choice trial sequences for all regimes, we compute the predictions of cognitively efficient agents for each problem. Since our model is parameter-free, we do not stratify problem exposure based on subject exposure to problems. However, comparative performance is only assessed based on problems for which corresponding subjects' choice behavior had been recorded.

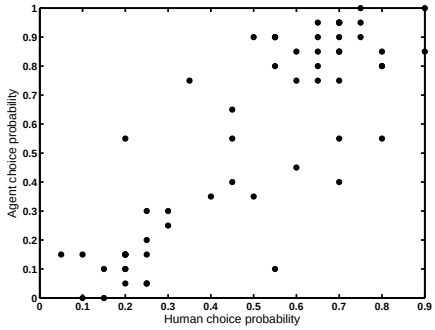
Results The results we obtain show (see Figure 4.6) that the behavior of cognitively efficient agents substantially resembles that of human subjects in the prospect risk equivalence task, under all three experimental conditions.



(a) Description (d)



(b) Experience with repetition (r)



(c) Experience with sampling (s)

Figure 4.6: Scatter plots visually showing agreement between aggregated choice behavior of subjects under all three conditions of the Erev-Roth risk preference experiment and predicted choice behavior of cognitively efficient agents in the same three settings. We see a significant degree of correlation in all three conditions(d = 0.68, r = 0.75, s = 0.86) without having to resort to statistical parameter tuning, which leads us to believe that our model captures deep aspects of human risk assessment

To better understand the significance of our results, observe that our model's performance of $\rho = 0.68$ in the description condition fares poorly against the competition's best result of $\rho = 0.92$ for a logistic regression model. For the repeated trials condition, our correlation value of $\rho = 0.75$ compares quite favorably with the best result obtained by an ACT-R model with sequential dependencies and blending memory. Finally, in the sampling condition, our score of $\rho = 0.86$ compares very favorably with the winning score of $\rho = 0.92$ obtained by a linear combination of four decision heuristics. While the raw empirical data suggests that our model's predictive performance is middling at best, such an assessment fails to account for the facts that (i) ours is the only model that can even make predictions in all three regimes and (ii) our model's results are reported *without* any parameter fitting. Fitting parameters such as working memory size, confidence ranges etc. can easily improve our empirical performance above its already creditable standards, but would not add any explanatory power to this exercise.

The fact that our model, which takes no account of the value of the lotteries, is applied without changes across all three experimental settings, and uses no statistical fitting to improve its fit with the data, performs as well as it does across all three conditions should be considered powerful evidence for its validity. To the best of our knowledge, ours is the first theoretical model to make predictions compatible with behaviors seen in both decisions from experience and decisions from description. We note, however, that a unification of the repeated trials and sampling regimes has already been proposed in [72].

4.4.5 Modeling reward-inference

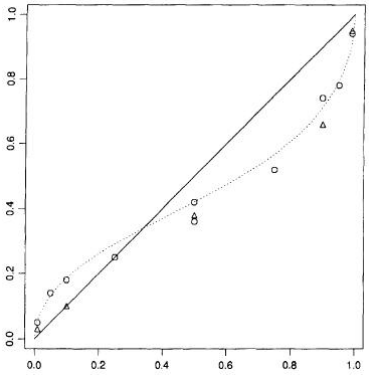
It must be acknowledged that our model's dependence on a specific definition of a reward-inference process imposes a significant limitation on its general applicability. While simply using binary relations of the form 'this outcome is preferred over that outcome' encoded as $\{0, 1\}$ and $\{1, 0\}$ is sufficient to demonstrate biases in probability etc in our experiments, adapting this model to settings where the environment demands a more structured representation will naturally require a more sophisticated model of how the agent orients itself with respect to the environment. At the same time, assuming a trivial reward-inference mechanism allows us to clearly distinguish the explanatory value of assuming information-theoretic coding of beliefs in memory.

Furthermore, adding incrementally sophisticated models of reward-inference to our cognitive model should result in falsifiable predictions about which cognitive biases we would expect to see in creatures capable of different levels of value representation. To take just one simple example, assuming our agents to simply be able to pick the better of two options makes our risk aversion results symmetric to gains and losses, which is incommensurate with the loss aversion documented in [4]. However, allowing agents to be able to quantitatively assess ratios between the risky option R and safe option S using a simple model (see Table 4.1) leads to the replication of loss averse behavior in our simulated agents, as shown in Fig 5.2(a). This leads to a testable prediction that animals capable of only comparing preferences but not assessing ratios, should demonstrate risk averse behavior, but not loss averse behavior. In measurable terms, assuming prevalent neo-Piagetian information processing models [103] to hold, this constitutes a human developmental prediction that in the economic sense of [32], human children will demonstrate risk averse behavior at an earlier age than loss averse behavior.

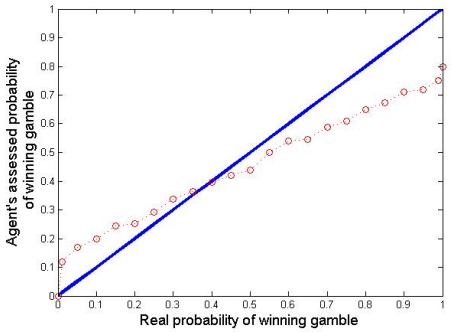
Table 4.1: Reward-inference for prospect theory experiment

	g(Risky)	**g(Safe)**
$R > S$	1 - S/R	S/R
$R < S$	R/S	1 - R/S

We assert that since different forms of value representation will require reward-inference models of different levels of sophistication, any detailed implementation thereof naturally lies outside the scope of the present paper. We further emphasize, in the light of the example we present above, that while differentiating our cognitive model from a specific reward-inference model does weaken its generality of predictive engineering applications, it also presents as an interesting direction for future scientific work, an opportunity to falsifiably test the predictions of different reward-inference models against experimental results in both the animal and human cognition literature.



(a) Loss aversion and risk aversion seen in human subjects in Kahneman and Tverksy's experiments



(b) Results from simulation of prospect theory experiment using cognitively efficient agents as subjects. Note the vertical asymmetry replicated.

Figure 4.7: 'Numerate' self-motivated learners show loss aversion, as opposed to their more primitive cousins in Fig 4.3(b). In a binary outcome space, as used in these experiments, the vertical asymmetry implies loss aversion since gains with a low probability are not as critical to subjects as gains with high probability, the latter of which are equivalent to losses with low probability

4.5 Discussion

4.5.1 A model for self-motivated reinforcement learning

Early efforts at using associative learning to model cognition took a strongly behaviorist stance(see e.g. [104]) with respect to their causal interpretations, a view which is no longer considered viable. In light of further advances made in cognitive science and machine learning, interest has recently re-emerged [20] in constructing RL-driven explanations for cognition. Current research, however, tends to restrict the purview of RL to providing mechanistic explanations at the conceptual level of neuro-physiological circuitry [105], with other more sophisticated theories invoked to explain higher cognition.

Through the fact that our algorithm basically gives an agent a stochastic sense for preferring some outcomes over others, as it continues to learn this sense in a sequential manner, it is structurally similar to reinforcement learning. Identifying quality-beliefs with policies leads to an immediate understanding of our choice-learning algorithm as a

model-free, on-policy reinforcement learner. There are, though, important differences, which are best explained as being threefold - differences in semantics, in epistemology and in teleology.

The semantics of our approach differ from traditional reinforcement learning in that, instead of representing the environment as a set of static states, with a menu of actions available to the agent, as is the case in RL, in our approach it represents the environment simply as a set of outcomes, thereby collapsing, in a subtle sense, the difference between the external and the internal world for the agent. This insight is best motivated with the 'embodied cognition' approach proposed in the AI community by among others, Hubert Dreyfus [106] and Rodney Brooks [107]. The embodied cognition view in AI suggests that, rather than modeling artificial agents as representing the world as an objective *out-there* object, it is better to represent it entirely in terms of the agent's interaction with it, thus making, as Brooks phrases it, 'the world its own model'.

The second change we make to the reinforcement learning paradigm concerns the epistemology of the learning agent. It is striking to note that, in translating the idea of utility from economics to computational learning, one of its fundamental aspects, the idea that utilities can only be coherently expressed as relative quantities, has been ignored. Recall that the original axiomatization of utility functions by [34] was motivated by the realization, following concerted psychophysics attempts to consistently estimate stimulus-response in humans, that a general model of absolute utilities could not be obtained. The seminal contribution of [34] was to work around the inaccessbility of absolute utilities by showing that, given four mathematical assumptions about the nature of utilities, viz. completeness, independence, transitivity and continuity, it was possible to construct an ordinal set of preferences that would not be inconsistent from the point of view of choice selection. That is, even though absolute values were not computable, a mutually consistent comparative ranking of options could still be obtained. However, in computational learning, the existence of absolute utilities is freely assumed [8]. While doing so is justifiable in artificial domains, where the value of a particular outcome can be encoded by a designer, any attempts to transfer such sequential machine learning approaches to model human learning must necessarily modify this assumption. We do so by considering agent beliefs to be assigned comparatively to various outcomes encountered, with the quantitative aspects of belief formation dependent on a domain-specific

reward-inference process.

The third change we make to existing reinforcement learning assumptions concerns the goal of the learning agent. Reinforcement learning was originally developed as a method for learning controllers in designed systems. The stability properties of these systems were expressed in terms that made maximizing long-term reward a natural statement of the required solution. However, in seeking to adapt the RL framework to human decision-making, a more biologically realistic goal was found to be essential. We have suggested that the goal of a biologically realistic decision-making agent is to learn sufficiently predictive theories with minimal cognitive effort.

We now further detail the evolutionary argument, mentioned earlier briefly, that justifies the goal we define as natural for biological agents. Consider a biological organism that is capable of observing its own preferences with respect to the environment[8] , but which needs access to resources in the environment in order to retain homeostasis. Assuming that resource availability fluctuates in both space and time, satisfactory communication with the environment would effectively become a prediction task, with the evolutionary goal being constructing theories of the environment sufficiently predictive to secure enough resources to ensure survival of the genotype. Furthermore, selection pressures may be expected to ensure that efficiency in the use of limited cognitive resources would be promoted in a population of such agents. It is in the light of this understanding of metacognitive intelligence as having evolved as a fundamentally predictive organ that we suggest that minimizing cognitive effort in constructing predictively adequate beliefs about the environment is a reasonable goal for humans in particular, and all metacognitive organisms in general.

By developing a cognitive model of learning using the mathematical machinery of reinforcement learning, our work opens up prospects for modeling higher cognitive processes in a new manner. While reinforcement learning has been used to model aspects of cognition before [109], our approach differs fundamentally from existing models in defining the utility and hence rationality of decisions in terms of an agent's internal cognitive processes. This *introverted* reinforcement learning approach, as we show in our results, leads our algorithm to display an array of sophisticated behaviors statistically incompatible with simplistic external-reward-averaging. Interestingly, a recent

[8] In other words, a metacognitive [108], or self-aware organism

effort similar in spirit to our own may be seen in [110], where the authors show how estimates of summary statistics, interpreted as a subjective prior of the controllability of the environment can be used to influence an RL agent's exploration and policy-learning ability, which they use to motivate a control-based theory of the development of learned helplessness. Here, as in our work, the decision variable of interest - control - is a psychologically meaningful abstraction, quantified by information-theoretic means. To the best of our knowledge, this remains the only other attempt at using reinforcement learning to explicitly model higher cognition by quantifying heretofore qualitatively understood cognitive phenomena.

While a fuller and more rigorous treatment of this matter remains an avenue for future work, we note in passing that it is fairly straightforward to map Huys and Dayan's composite notion of control as confidence in our setting. The link with their entropy measure is trivially seen, the link with the achievability of outcomes is obtained by the embodied-outcome based representation that we employ, and the link with reward achievability, with a key modification, is obtained via the history of surprise informing an agent's preferences. Our concept of confidence therefore, resembles Huys and Dayan's concept control except in that, instead of assuming that agent's preferences for outcomes are governed by which one corresponds to greater absolute reward, we assume this notion of goodness of outcomes to be relative.

4.5.2 An information-theoretic model of memory

The principal technical novelty in our decision-making approach lies in our replacement of the statistical notion of descriptive complexity with one of cognitive processing cost in an information-theoretic model selection setting. Establishing a plausible definition of cognitive processing cost has been a major open question in both the AI and cognitive science communities [91, 92]. In our modeling effort, we realized that focusing entirely on modeling cognitive costs (holding perceptual costs of decision-making constant) allows us to pose the larger processing cost problem as a memory access cost problem. We construct a model of memory by hypothesizing that memory storage follows an information-theoretic optimal coding principle sensitive to the predictive value of the stored belief. In a novel modification to existing ideas about intrinsic motivation (which simply assign higher motivation when predictions fail [111, 5]), our use of

an information-theoretic criterion implies that both extremely predictive beliefs and extremely unpredictive beliefs will be exceptional. Assuming random access to the memory, a natural definition of memory access costs follows.

Several existing cognitive architectures use a model of memory similar to the one we have presented. Agents select a subset of prior beliefs to populate active memory. In existing models of memory, e.g. ACT-R [112], Soar [67] etc, the appropriateness of a prior belief in a context is determined in the form of a similarity between this prior belief and the agent's query/current belief. It is easy to see that this is structurally similar to our approach. The difference lies in our replacement of a similarity measure on the set of declared attributes of a decision instance (as in ACT-R) with an informational exceptionality measure on actual beliefs imputed to a decision instance. In view of the surprising novelty of our existing results, it seems to be a simple and worthwhile task to incorporate our informational exceptionality metric in these architectures to test its validity on decision domains of greater representational complexity than our simple experiments can hope to achieve in the near future. Finally, we note that, notwithstanding structural similarities, a direct performance comparison of our memory model with earlier models is not trivial, since we model performance in terms of cost of memory access, while these models measure performance in terms of errors in access. While it is easy to see that these two measures are likely related[9] , and while efforts to model this relationship have been made [24, 113], we believe mapping the quantitative implications of our model to the production model view of memory will require some further research.

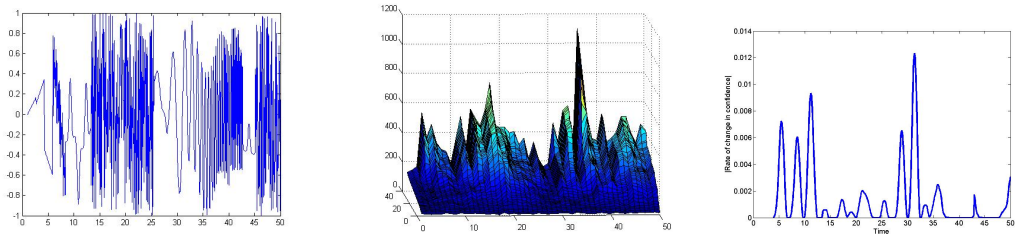
4.5.3 A causal model of intrinsic motivation

Our information-theoretic formulation of memory coding emergently provides an alternative account of intrinsic motivation. Until fairly recently, the idea of using statistical information about the environment to determine intrinsic motivation had only been qualitatively addressed [114]. The observation that the activity of dopamine neurons could be modeled using temporal difference reinforcement learning methods [26] has led to proposals of intrinsic motivation based on the magnitude of the error in predicting

[9] Beliefs that are too costly to access would be expected to be more likely to encounted errors in access

expected rewards (see e.g. [111]). However, it has been more recently demonstrated that the neural substrates in question appear to code not for reward prediction errors, but prediction errors in general [115]. This observation suggests that a more sophisticated approach to encoding motivation is needed to encompass motivation unrelated to rewarding events.

Following Czikzenthmihalyi's proposal [116] that agents are motivated by an intrinsic 'curiosity' to search for situations with an intermediate degree of 'novelty', Oudeyer and colleagues have developed a model of intrinsic motivation [117] called *intelligent adaptive curiosity* (IAC). The IAC model of motivation proposes that agents choose to explore 'regions' in the environment where they expect to make the most learning progress. Learning progress in turn is quantified in terms of the reduction in mean prediction error. Thus, an IAC agent essentially selects outcomes that seem predictable (manifest as strong error reduction rate) but not too predictable (where the error rate reduction would be stagnant).



(a) Sample environment with varying function complexity

(b) Total exploration over time with R-IAC algorithm

(c) Rate of change in confidence for self-motivated learning

Figure 4.8: Results for complexity affinity experiments showing our model evincing greatest learning ability (measured via rate of change in confidence) during intervals of intermediate complexity, as predicted for biologically realistic motivation systems by Oudeyer et al.

Relatedly, in a reinforcement learning setting Simsek et al [118] have proposed a model of internal reward/motivation that tries to ensure efficient exploration of the environmental state space. The intrinsic reward, in this case, is determined via the amount of change in the expected utility of *policies* associated with different states. Both the Simsek and the Oudeyer computational models of motivation share an underlying expectation with our model of learning in desiring to improve the ability to

learn, and seeking to associate intrinsic motivation with an agent's self-perception of its predictive ability. However, by disregarding the cognitive machinery involved in the process of learning and tying the intrinsic motivation directly to the statistically measured learning rate, the Simsek approach fails to replicate the bistable nature of natural agents' motivations, viz. they stop trying to learn both when the environment becomes too predictable and when it becomes too unpredictable, an insight that the Oudeyer model captures.

The Oudeyer model, in turn, however, imposes this bistability from the intuition that agents seek out experiences with intermediate novelty, which, on the surface, directly contradicts our model which assigns greater significance to both extremely unpredictable and extremely predictable events. However, it it important to remember that while Oudeyer et al are directly trying to model motivation, our model is one of memory access, which operationally creates a model of motivation.

Further, a closer examination of our model reveals that situations with high predictability, if sustained, cause the agent to stop updating its beliefs, since no further increases in confidence are possible once certainty is assured. On the other hand, highly unpredictable decision instances lead to low confidence, preventing further updates of similarly confidence-reducing novel instances. The bistability of motivation with respect to event predictability is hereby retrieved. The recovery of the basic qualitative outlines of the IAC framework of curiosity from our information-theoretic model of memory access is unlikely to be purely coincidental. While a deeper understanding of the connections between IAC and our model of learning lies outside the scope of this paper, it constitutes a very interesting prospect for future research.

We note finally that novelty and predictability are not the same concept. A highly predictable event (hearing a child speak for the first time) can be extremely novel, while a highly unpredictable event (a coin toss) may have very little novelty. Thus, the mapping between novelty and predictability made in the IAC literature is problematic from the perspective of cognition. Our model, on the other hand retrieves the prediction-novelty relationship in a more cognitively meaningful way. A predictable event of which an agent has no past contextual memory will affect the agent's beliefs and confidence (internal reward) far more than a similar event for which the agent has a long history of prediction. In contrast, both novel and expected unpredictable events will be preferentially ignored

by the agent. Within a specific context, therefore, an introverted agent, therefore, will learn best either in situations that are completely novel with an agreeable degree of predictability, where any additions to its belief store can only increase its confidence, or in familiar situations with a degree of predictability higher than it has erstwhile known. We believe that the first case is explained in IAC, whereas the second, involving increased sense of control, is not.

4.5.4 A new basis for judging rationality

In the classical expected utility paradigm, an agent's reward is identified as some analytically tractable mapping from an objective numeric value obtained from the environment. The goal of the agent is presumed to be to maximize the reward it can obtain from the environment. As we have briefly detailed earlier in this paper, this effort at modeling decision-making, while quite useful in some specific artificial settings has proved to be a poor model of realistic human decision-making.

In recent years, there has been a shift away from trying to construct causal models of decision-making towards simply finding heuristics that have predictive value within particular domains of interest. To a very large degree, this shift in emphasis has resulted from the appearance of a large number of deviations from the predictions of expected utility theory in diverse domains of application. The presumed analytical intractability of this menagerie of cognitive biases has led to a pessimism towards finding unified models of behavior and decision-making, causing the conjectured view of humans to shift from being considered rational utility maximizers to being considered evolved 'adaptation executors' [119], thereby opening the door to a wide array of specialized heuristics with no ontological interpretations to take their place as the state-of-the-art in decision theory.

In this work, we have showed how three different families of cognitive biases can, in fact, be generated from a single causal model of decision-making, merely by shifting the objective of a classical bounded rational agent from resource-constrained utility maximization to prediction-constrained cognitive effort minimization. By doing so, we have essentially proposed a new way of defining rational utility, which subsumes positive aspects of both the classical expected utility paradigm [34] and more recent heuristic-based

methods [85] while avoiding their defects. Specifically, our model retains the analytical tractability and causal interpretability of the traditional expected utility/rational choice paradigm while adapting the definition of rationality to confirm with Gigerenzer's [1] idea of 'ecological rationality'. By adopting an embodied representation of the agent-environment interface, and an information-theoretic basis for defining costs, we are able, however, to generalize our model's dependence across the ecology of different domains. Thus, we avoid having to conjecture multiple models for different domains of decision-making; our models remain ecologically rational across multiple contexts by remaining information-theoretically rational.

While our model is rational in the strict economic sense of the term, a few other specialized definitions of rationality in the context of intelligent agents have been proposed. Somewhat surprisingly, we find our approach to adequately satisfy a considerable number of these definitions. A comparison with Anderson's [120] 'rational analysis' requirements shows that our model, notwithstanding its evolutionary and information-theoretic motivations, can be firmly grounded as a rational model of human cognition. We satisfy all six of Anderson's requirements:

1. Goals. We specify a well-defined evolutionarily motivated goal.

2. Environment. We define a specific form of engagement with the environment, in the form of our reward-inference update.

3. Computational limitations. Computational limitations form the crux of our modeling effort; we characterize them quantitatively using our definition of processing cost, which in turn depends on our particular model of memory access.

4. Optimization. We provide a closed-form objective function, and an algorithm that optimizes it.

5. Data. While some current efforts at modeling intrinsically motivated reinforcement learners attempt to show human-like behavior on simulated toy examples [121, 122], we validate our model against results observed in human subjects and find good confirmity with empirical evidence.

6. Iterate. Agents following our model will iteratively refine their quality-beliefs to better serve their predictive purposes.

More recently, Chater [42] has differentiated 'mechanistic' reinforcement learning models that posit the existence of specific neuro-physiological machinery for their implementation from 'rational' models that suggest that reinforcement learning strategies are merely instantiations of a more general cognitive mechanism in cases requiring information processing in ways where such strategies are known to be optimal. He further proposes that the experimental evidence on various tasks supports the validity of the latter class of models. It is pleasant, therefore, to observe that while the present instantiation of our model on a sequential decision-making task takes the contours of a reinforcement learning algorithm with memory resampling, the basic principle of cognitive cost minimization under prediction quality constraints underlying it is substantially more general and could be used to construct models of learning and behavior under other task representations. Our model is therefore, 'rational' by Chater's definition.

Finally, we recall that Mill's original definition of utilitarianism [2] states simply that it involves getting the most reward for the least effort, without specifying which quantity is to be optimized and which to remain satisfactory. Traditional rational choice models have chosen to maximize external reward, bounded rationality models [123] have chosen to maximize external reward while bounding internal cognitive costs. Our approach simply inverts this objective by optimizing internal cognitive costs while bounding external reward via the predictive potential to acquire it. Thus, even though our model might appear radical at first glance, it, in fact, is a dual of standard bounded rationality interpretations of Mill's original definition of utilitarianism.

While the satisfaction of prior qualitatively expressed notions of rationality does not add empirical support to our theory, it should serve as an indicator of the basic *reasonableness* of our approach, which is an important criterion when evaluating a model fundamentally premised on optimizing a variable that is not directly observable. It may well emerge that the practical applicability of our model will be severely hamstrung by the dependence on revealed preferences to infer cognitive processing costs. In such an eventuality, we would still not count our labor fruitless, as it would at least have outlined what a rational theory of decision-making for self-aware agents *could* look like.

4.6 Conclusion

We have constructed the first coherent quantitative model for self-aware learning and decision-making. In doing so, we have proposed a novel information-theoretically motivated basis for quantifying cognitive processing costs. We find that our model explains empirical data corresponding to observed cognitive biases much better than existing models of decision-making, and does so with an amazing level of generality across datasets and experimental settings. The surprisingly good predictions from our model lead us to suspect that we may have uncovered a strong information optimization heuristic evolutionarily embedded in the nature of human cognition.

Chapter 5

Realistic goal-directed learning

5.1 Introduction

" (Reinforcement learning) proposes that whatever the details of the sensory, memory, and control apparatus, and whatever objective one is trying to achieve, any problem of learning goal-directed behavior can be reduced to three signals passing back and forth between an agent and its environment: one signal to represent the choices made by the agent (the actions), one signal to represent the basis on which the choices are made (the states), and one signal to define the agent's goal (the rewards)" [8].

While this abstraction is useful in handling agent-environment interactions in artificial domains, any effort to construct models of higher-level cognition in biological organisms based on reinforcement learning must grapple with precisely the sensory, cognitive and control apparatus that its existing theoretical structure seeks to elide. Because of its control theory antecedents, reinforcement learning makes unrealistic assumptions about the epistemological abilities of biological organisms.

Specifically, it assumes the existence of cardinal and state-specific 'reward' values are provided to the agent by its external environment. This focus on reward being computed in the environment has traditionally been explained as being a semantic issue, which can be resolved by placing the agent's internal reward generation apparatus on the outside of the arbitrarily defined agent-environment boundary. However, such an explanation cannot account for the internal cognitive costs of the agent's decision-making process itself. In other words, while agents' energetic and situational costs can

be placed in the environment, it is not clear if the same can be done for entropic or decision costs inherent in the process of action selection itself. As a consequence, any model of learning built using regular RL is forced to predict future agent behavior strictly as a function of environmental inputs. Such a strong behaviorist stance is incompatible with empirical data and is considered unrealistic [42]. Hence, it appears essential that a generalization of reinforcement learning towards realistic goal-directed learning take mentalistic processes into account via an account of intrinsic motivation emerging from the cognitive apparatus of typical biological agents.

There have been sporadic attempts in the AI community in recent years to devise decision-making agents that are motivated by self-preservation, novelty or other qualities viewed as *intrinsic* to biological agents' understanding of their environment. A number of these attempts e.g. [111, 118, 121, 6] have used the formal structure of reinforcement learning (henceforth RL) to obtain models of decision-making and control that better reflect the capabilities of realistic intelligent agents. The convergence of these efforts seems to follow from an intuitively obvious, but as yet formally elusive, mapping between the reinforcement learning formalism and real-world decision making. It appears natural to attempt to unify these approaches into a general theory of higher-level cognition and choice selection grounded in reinforcement learning.

It is now generally recognized that the choice-selection behavior of intelligent organisms depends fundamentally on their inferring value of outcomes insofar as these outcomes satisfy their intrinsic biological and psychological needs [124, 122, 16]. This observation immediately leads to a more nuanced understanding of value appraisal as indirect and intrinsic *reward-inference* in service of biological needs. Furthermore, the value of a particular option can only be appraised by an agent relative to the value of other possible options in a decision context, rendering the notion of absolute cardinal reward meaningless. In its place, a realistic understanding of value appraisal predicates viewing rewards as emerging intrinsically through an agent's appraisal of the relative desirability of multiple possible outcomes. Reward relativity is by no means a radical assumption. Seminal work by [34] was based on the idea that a consistent relative ordering over the value of different options could be obtained, leading to the establishment of the von Neumann-Morgenstern (VNM) expected utility axioms that yield mathematical

conditions under which relative order can be represented by a cardinal utility representation. However, the VNM axioms are well-known [84] to be violated in practice and do not have empirical support as realistic principles describing organisms' subjective sense of value. Representation relativity, therefore, simply constitutes an epistemological retreat to a pre-VNM state of understanding of value as being represented by not necessarily consistent ordinal preference orderings.

This understanding of realistic reward representation causes us to reformulate the optimal choice selection problem under uncertainty in a novel and fundamental way. In light of the observation (supported independently by [16, 122, 5]) that realistic rewards are fundamentally intrinsic, it immediately follows that such intrinsic rewards are to be measured, not with reference to some absolute zero point corresponding to a Platonic 'no reward' setting, but with respect to the agent's current preference with respect to various environmental outcomes. In other words, intrinsic rewards must be computed not with respect to outcomes, but with respect to policies and hence relative to agents' existing hedonic set points.

Finally, a realistic sense of valuation must emerge entirely from the selection of options that promote the existential or survival goals of organisms. Removing the locus of optimization from external reward to intrinsic valuations forces us to rethink our sense of optimal behavior in a fundamental way. If the external goal of agents is continuing to adapt to dynamic environmental conditions, the only possible optimization would be one that allows them to do so at minimal internal cognitive cost. Thus, a realistic model of choice selection must specify both the effect of an agent's understanding of the desirability of its environmental options and the effect of internal cognitive processing costs on the agent's process of understanding itself. In this paper, we show how investigating this line of thinking leads us to a more realistic formal model of choice learning. We then demonstrate a solution to the choice learning problem posed in our framework and show that its behavior reflects interesting properties seen in human subjects but not replicated in existing choice models. Finally, we describe how our framework generalizes both standard RL and other contemporary efforts at creating intrinsically motivated agents.

5.2 From reinforcement learning to realistic learning

Let $x \in \mathcal{X}$ correspond to the agent's internal state representation of the environment. For purposes of simplicity, we embed the notion of state transitions into states in this paper, i.e., $x'|x \rightarrow z$. This follows from the technical observation that state transitions can be tabulated just as easily as individual states and the intuition that natural agents typically are concerned with *process* more than structure, causing state transitions to be natural objects to track. Given this understanding, we define $u(z)$ as a probability distribution over state transitions that encodes the effects of the controller's actions. Furthermore, in light of our novel understanding of state desirability, as described in Section 5.1, being a relative comparison between different states, we also represent the agent's relative preferences as a distribution across available states, $p(x)$. The second novel aspect of our formulation is to ground the agents relative preference $p(x)$ in an inferential process. We assume that the agent receives information about the relative intrinsic value of options $p_k(x)$[1] at decision k, which can be expressed in terms of a distribution representing the inferred relative worth of available options. Given this new information, the agent updates its understanding of a distribution across preferences $P(p(x)|m)$, where m is a parameter that controls the complexity of this distribution. Since $P(p(x)|m)$ stores all the preference information from the agent's history of engagement with its environment, we call this a memory distribution. In addition, real world environments are dynamic and tend to cycle between modes, e.g., chasing prey, resting under tree, drinking water, which creates an implicit dynamics $P(p'|p, m)$. We take a filtering approach to tracking the distribution across $p(x)$, described below. It is important to note that the concept of reward is fundamentally revised in our outlook. Rather than being grounded in environmental states, we view intrinsic reward as the agent's assessment of the current decision is light of the factors that affect the decision making process, including the cognitive costs of making the decision. In our model, two factors are critical for determining intrinsic reward: the cognitive cost of obtaining a preference about the relative goodness of different options and the predictive or functional value provided by using this preference to guide future actions. We consider the functional $T(p, x)$ to encode memory access costs required for preference formation. We

[1] We suppress the k subscript where not needed. p may be understood to be p_k for arbitrary k unless explicitly stated.

use a KL divergence [69] to measure the atypicality of a preference $C(p, p^*)$ in order to estimate the extent to which a particular preference p differs from the agent's typical sense p^* for the goodness of options in a particular domain. Preferences that are vague or distributed uniformly across multiple possible outcomes are not useful, causing us to include a perceived uncertainty cost $H(p)$. For our present purposes, we simply take this to be the Shannon entropy of p. Preferences that are both typical and specific are considered useful. Finally, agents' action selection may be constrained through partial domain controllability. We include a control cost $R(p, u) = KL(p, u)$, which penalizes deviations in control $u(x)$ away from the current preference $p(x)$, again measured using a KL divergence. Thus, our intrinsic reward function can be written as,

$$r_{intrinsic}(x, p, u) = T(x, p) + C(p, p^*) + R(p, u) + H(p). \tag{5.1}$$

Our agent minimizes this intrinsic reward function over an infinite event horizon with respect to both controls $u(x)$ and memory representation m:

$$V_{u,m} = E_{p,x}\left[\sum_{k=0}^{\inf} r_{intrinsic}(x, p, u)\right]. \tag{5.2}$$

Since our representation of this control problem requires us to introduce an additional memory distribution $P(p|m)$ on preferences about state desirability, it formally resembles the belief MDP formulation proposed by [125]. The key formal difference between standard belief MDPs and our approach is that our intrinsic reward function depends on the belief state, and that our agent's control problem includes controlling the memory distribution. Recall that the standard Bellman equation for belief MDPs is,

$$V^*(b, x) = \max_u \sum_{b', x'} P(b', x'|b, x, a)\left[r_{extrinsic}(x, u) + V_u(b', x')\right]. \tag{5.3}$$

Homologously, the most general Bellman equation for our problem can be written as,

$$V^*(p, x) = \min_{u,m} \sum_{p', x'} P(p', x'|p, x, u, m)\left[r_{intrinsic}(p, x, u) + V^*(p', x')\right]. \tag{5.4}$$

To develop a solution for (5.4) we must first obtain an expression for $P(p', x'|p, x, u, m)$, and then determine how to optimize the value function with respect to both the action

policy and the memory distribution. The transition probability $P(p', x', |p, u, x, m)$ can be factored as,

$$\begin{aligned} & P(x'|p, p, u, x, m)P(p'|p, u, x, m), \\ = \; & P(x'|u, x)P(p'|p, u, x, m) \qquad (5.5) \\ = \; & u(x)P(p'|m) \qquad (5.6) \end{aligned}$$

where the simplification results from the facts that (i) the state transition $P(x'|p, p, u, x, m)$ is independent of the preference $p(x)$, and thus completely determined by the policy $u(x)$, (ii) that the belief transition is outcome-invariant, i.e. $P(p'|p, m) = P(p'|m)$. The agent controls the u transition, as well as the parameter m that controls the p transition. This factorization leads to a simpler representation,

$$V^*(p, x) = \min_{u,m} \sum_{p',x'} u(x)P(p'|m) \left[r_{intrinsic}(p, x, u) + V^*(p', x') \right]. \qquad (5.7)$$

The Bellman equation (5.7) captures our modified assumptions about reinforcement learning and would be amenable to standard value iteration type solutions if we could compute the expectation under the memory distribution tractably. Rather than making parametric assumptions about the distribution $P(p'|m)$, we represent the distribution via a set of observed samples, so that $P(p'|m) \; \sum_k w(k)\delta(p - p_k)$, where $w(k)$ represent weights that represent the importance weight for that observation. By adjusting these importance weights, the agent can control its memory distribution. The simplest version of this control is when $w(k)$ are set to either zero or one, which intuitively corresponds to selecting a subset $\mathcal{M}'$ of memories $p \in \mathcal{M}$ that it has previously experienced. Thus we can interpret the control parameter m as a selection and composition process on memory 'particles'. This selection process acts like a particle filter approximation which allows the expectation across $P(p|m)$ to be tractably computed. Equation (5.7) then becomes a dual optimization problem,

$$V_{u,m}(x, p) = \min_{u,m} \left[\sum_{p_m \in \mathcal{M}'} \sum_{x \in \mathbf{x}} u(x)r(x, u, p_m) + \sum_{p_m \in \mathcal{M}'} \sum_{x \in \mathcal{X}} u(x')V_u(x', p_m) \right]. \qquad (5.8)$$

A general solution to (5.8) is non-trivial and is beyond the scope of this paper. Here, we consider an intuitive greedy approximation to the formal setup, involving only a

one-step lookahead. This simpler problem can be solved as two separate optimizations, one across u, and one across m. In particular, observe that the cost functional $r_{intrinsic}$ can be divided into two components, $Q(x,p) + R(p,u)$, with Q encoding preference formation costs with no action policy dependence and R encoding controllability costs as a KL divergence between u and p. Observing the summation across all $p \in \mathcal{M}'$, we can apply Jensen's inequality to obtain

$$KL\left(\frac{1}{|\mathcal{M}'|}\sum_{p_m \in \mathcal{M}'} p_m, u\right) \leq \frac{1}{|\mathcal{M}'|}\sum_{p_m \in \mathcal{M}'} KL(p_m, u),$$

since the KL divergence is convex in the first argument. The left hand side of this expression lower bounds R, and can hence be minimized as a surrogate. This construction also has a nice intuition of collapsing the evidence from multiple preferences into one composite preference, thereby naturally instantiating $p^*(x)$, which reflects the agent's recollection of its typical preferences. We therefore obtain optimality at

$$u^*(x) = p^*(x) = \frac{1}{|\mathcal{M}'|}\sum_{p_m \in \mathcal{M}'} p_m. \tag{5.9}$$

Since R contains the only u dependence for $r_{intrinsic}$, (5.9) is analytically the optimal solution for the greedy u optimization. Given optimal u, it now remains to optimize across m to complete our description of a greedy approximate solution for (5.8). However, the memory optimization procedure is embodied in the structure of the agent's cognitive memory apparatus. Thus, the m optimization must emerge from the description of a cognitively realistic memory model. We turn to this task next.

5.3 Realistic learning needs realistic memory

To optimize memory, we plug the optimal action u^* into (5.8), expanding the reward term using (5.1), to obtain the second minimization:

$$V_{u^*,m}(x,p) = \min_m \Bigg[\sum_{p_m \in \mathcal{M}'} \sum_{x \in \mathbf{x}} u^*(x)(T(p^*, p_m, x) + C(p^*, p_m) + R(u^*, p_m) + H(p_m)) + \sum_{p_m \in \mathcal{M}'} \sum_{x \in \mathcal{X}} u^*(x') V_u^*(x', p_m) \Bigg], \tag{5.10}$$

To instantiate T we must answer, 'what constitutes optimal memory selection?' Memory optimization should be grounded in the survival goals for biological agents, which principally revolve around energetic and entropic homeostasis[5]. The need to minimize uncertainty about the environment together with a finite bandwidth of cognitive processing resources suggests agents should minimize their cognitive processing costs to generate predictive models of the environment. We model cognitive processing costs directly as the cost of accessing previously observed preferences. We hypothesize that the access cost of the preference associated with a particular memory instance is determined by its predictive exceptionality, which in turn can be measured as a departure from the usual level of *surprise* that the agent experiences in making its predictions. Interestingly, recent experimental data provides support for such an information-theoretically optimal encoding existing in human subjects' memory [126]. The surprise S experienced by an agent operating with a preference p_a in comparison with a different preference p_b can be quantified with an information divergence of the form,

$$S(p_a, p_b) = \sum_{j=1}^{n_a} p_a^j \log \frac{p_a^j}{p_b^j}. \tag{5.11}$$

The predictive exceptionality of a past preference p' with respect to the current preference p (and hence the ease with which it will be available for recall to the agent) can then be defined as the deviation from the average surprise experienced by the agent $\overline{S}$ is

$$A(p, p') = |S(p, p') - \overline{S}|. \tag{5.12}$$

The cognitive processing cost of selecting a working memory $\mathcal{M}'$ from long-term memory $\mathcal{M}$ is simply the inverse exceptionality-weighted sum of the nominal cost of accessing all preferences p in $\mathcal{M}'$. Thus we define T as

$$T(p^*, p_m, x) = A^{-1}(p^*, p_m^{(x)}), \qquad p_m^{(x)} = \max\{p_m^{(x')}\}, \forall x' \in \mathcal{X}. \tag{5.13}$$

Observe that preferences are assumed to add a processing cost only to the outcome that they favor in probabilistic terms. Hence, over time, outcomes corresponding to less historically computationally expensive decisions become more preferable. Since the selective memory recall is also approximated by a greedy one-step look-ahead, we can regard the optimization over m as simply minimizing the immediate intrinsic reward

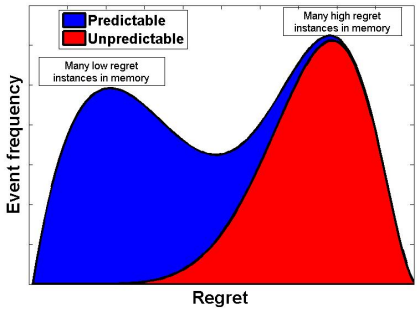
term. Note that in our particular implementation, the memory control parameter m simply takes on the interpretation as a selecting indicator vector over all preferences indexed in long-term memory. Thus, for its currently optimal action policy $u^*(x) = p^*(x)$, the agent recalls salient experiences from its memory by solving a further simplification of (5.10) of the form,

$$\underset{m}{\operatorname{argmin}} \left[\sum_{p_m \in \mathcal{M}'} (A^{-1}(x, p_m) + C(p^*, p_m) + R(u^*, p_m) + H(p_m)) \right], \tag{5.14}$$

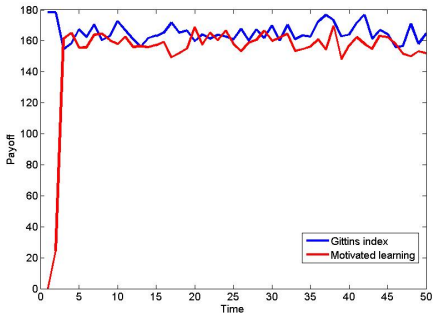
with the expectation under u^* ignored by virtue of being independent of m. Note that the contribution of the R term duplicates that of the C term in this expression, being simply $KL(p^*, p_m)$. However, this is a consequence of the greedy approach that sets $u = p$ and will not be true in general. The memory recall procedure we have defined reflects the intuition that an agent parsimoniously constructs its preference about its environment by recalling past preferences into its working memory in a way that minimizes its experienced cognitive processing cost while simultaneously reducing its environmental cost as well as the cost of preference representation with respect to the decision problem. Solving this optimization problem gives us the working memory subset $\mathcal{M}'$ and simultaneously a distribution over the agent's current preference about its options and an expectation of the intrinsic cost-to-go for various outcomes under the memory distribution. In practice, we solve this subset selection problem in a naive manner by sorting available preference samples by exceptionality and populating $\mathcal{M}'$ incrementally until the objective function converges, upper-bounded by a working memory size threshold. Continuing to solve for u and p iteratively allows the agent to navigate its environment intelligently by making useful predictions with minimal cognitive effort.

5.4 Experiments

In this section, we try to provide a functional intuition into how our reward redefinition practically affects an agent's sense of optimality by running simple simulations. In general, we find that our model behaves like a traditional reinforcement learner in domains where statistically typical options are the most rewarding, but diverges in interesting ways from classical predictions in domains where this property does not



(a) Diagram showing how an agent's memory recall will be influenced by the relative (un)predictability of the goodness of its options.



(b) The performance of our model tracks optimal Bayesian performance for standard multi-arm bandit problems.

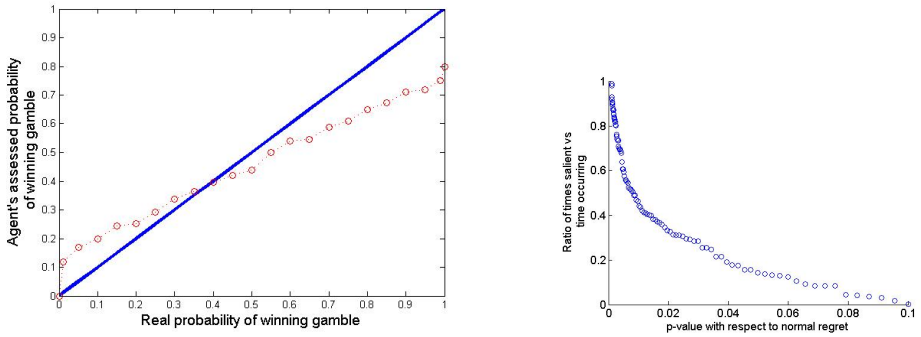
Figure 5.1: Agents behaving according to our model resemble classical reinforcement learning strategies in settings where statistically typical outcomes are also the most desirable, as is the case in most artificial control problems. This allows a statistical expectation to capture the information necessary to behave intelligently in the domain.

hold. Figure 5.1(a) shows a stylized view of the relative frequency distribution of low and high regret decision instances in predictable and unpredictable domains. It is important to realize that predictable domains will have a number of statistically typical events that the agent will learn to predict, thereby experiencing low regret many times. On the other hand, in unpredictable domains, there will be no typical events, resulting in a large number of high regret memories. In the former case, low regret instances will dominate memory recall, leading to prediction of statistically typical outcomes which improve both $T(p^*, p)$ and $C(p^*, p)$. In the latter case, the agent will struggle to select between a number of preferentially distinct high regret memories that improve $T(p^*, p)$ by possessing high exceptionality but reduce $C(p^*, p)$ since they are statistically atypical. Agents' selections in such domains may appear idiosyncratic and irrational, since statistically atypical options are chosen, and once selected, will persist.

Figure 5.1(b) shows that our algorithm performs close to optimality in the standard multi-arm bandit setting, where the Bayesian Gittins index solution is known to be both optimal and analytically tractable. The reward on each arm is Gaussian distributed with different means but known variance. Results are reported by averaging over 10 trials of 50 time steps each. Because our algorithm is implemented using a greedy update with

no look ahead, it can't solve the exploration-exploitation problem so we implemented an ϵ-greedy exploration strategy with $\epsilon = 0.05$. This shows that our model can be rational in the traditional sense: with enough experience it selects options associated with more rewarding cues.

In Figure 5.2(a), we show how, in more natural domains, such as the certainty-equivalence experiments of [32], our model generatively replicates behavioral results from classic studies resulting in the birth of prospect theory. In order to simulate the experimental setup described in [32], we design our outcome space to consist of two possible outcomes: select safe prospect or select risky prospect. For every decision instance, the payoff for the risky prospect is sampled from a Bernoulli distribution appropriate for the gamble. For the gamble in the example above, this means that the risky prospect will pay \$0 in about 9 out of every 10 decision instances. The reward-inference signal is constructed to assign a preference of 1 to the better prospect (and 0 to the worse prospect) at every instantiation. Thus, a choice between a gamble with a 0.1 probability of paying off against a certain safe outcome is modeled as a generative mechanism for reward-inference that reflects a selection [0 1] biased towards the safe choice 90% of the time and the alternate risky choice [1 0] 10% of the time.



(a) Results from prospect theory experiment using agents guided by our model as subjects.

(b) Peak experience dominance in sequential choice experiments

Figure 5.2: Self-motivated learners show both risk and loss aversion in accord with prospect theory predictions. Our particular memory model formulation also leads to peak-end effects [102], wherein agents over-represent rare preferences that correspond to peak experiences in memory recall.

We provided each one of a population of 200 agents with a series of 100 such reward-inference signals. A series is presumed to indicate the 'learning' phase for an agent with respect to a particular choice problem involving risk evaluation. At the end of a series, the agent is assumed to possess, in the form of its final preference, an evaluative model for selecting between the prospects offered in the [32] selection task. We modify the probability of winning or losing the gamble by modifying the Bernoulli distribution parameterizing the reward inference distribution.

Furthermore, our algorithm's behavior on this experimental task supports the observation [102] that the value of past experiences is assigned largely through evaluating them at their 'peaks', not in terms of an average over the entire experience (see Figure 5.2(b)). In our case, the peaks are determined, not in terms of absolute reward, but in terms of surprise experienced by the agent. Instances corresponding to extremely low or extremely high surprise predominantly influence our existential agent's internal preferences with respect to a sequence of events.

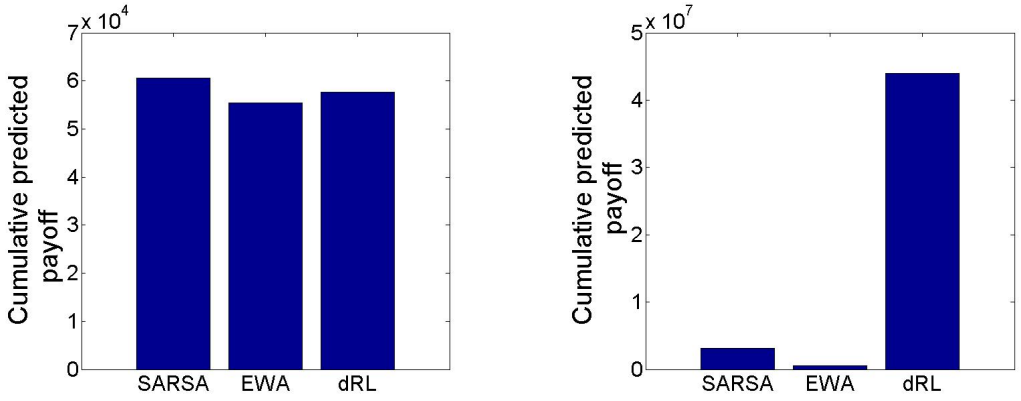
Finally, we ran simple experiments to illustrate the value of our algorithm's sensitivity to extreme risk events. Specifically, we show that cognitively efficient learning, in contrast with SARSA [127] and EWA [9], over-weights 'extreme' low-probability events, which can lead to more productive predictions in domains that demonstrate a heavy-tailed reward distribution. In our simulation study, the outcome domain is modeled as a simple grid world task. Reward distributions are probabilistically assigned to arms, and the parameters of the reward distributions are sampled from priors with fixed hyper-parameters.

Reward locations for a particular decision instance are selected by Bernoulli trials on each outcome. The Bernoulli parameter thus governs the sparsity of reward locations. For all locations selected as active, reward distribution parameters are sampled from priors and a reward instance is sampled from the reward distribution corresponding to the sampled parameters. The selection of active reward sites and the reward distribution parameters assigned to each grid locations stays persistent across multiple decision instances until/unless the grid is reset, at which point the process is repeated. Individual rewards are randomly generated for every decision instance.

All three algorithms are implemented in MATLAB in their basic forms [127, 9]. The state exploration of each algorithm is generated randomly from all N^2 outcomes.

We did not optimize the SARSA or EWA parameters for each environment, since our conclusions should hold for all parameter settings.

We measured each algorithm's performance on this grid world in terms of the expected future payoff with respect to the algorithm's current preference. For SARSA, we transform the Q-matrix by normalizing the values across options. To make SARSA competitive in learning time, we also enforce that preferences on options are start-state independent. For EWA and cognitively efficient or goal-directed learning (dRL), the preference relation was directly computed from the algorithm. The predicted payoff was summed over multiple decision instances to arrive at a cumulative predicted payoff for a run. This quantity should indicate the quality of the algorithms' performance in the multi-stage decision process.



(a) Predicted payoffs from normally distributed rewards

(b) Predicted payoffs from Pareto distributed rewards

Figure 5.3: Results from 1000 runs of N-stage decisions (N=100) in the grid world simulation. Quantity measured is the cumulative predicted reward given preference at step n and reward at step $n+1$ summed over all stages. Results from all runs are averaged. dRL performs comparably with SARSA and EWA when rewards are generated from normal distributions but outperforms them both when rewards are drawn from Pareto distributions.

For normally distributed rewards, our algorithm's performance is lower than SARSA but is comparable to EWA (Fig 5.3(a)). This intuitively follows, since the cognitively efficient algorithm persistently uses a small set of salient decision instances rather than all possible outcomes. Nevertheless, this satisficing approach is close to SARSA-optimal,

Table 5.1: For 100 trials each containing 100 runs, each in turn containing 100 stages, we tabulate the number of trials in which each of the three competing algorithms outperform the other two for both normally and Pareto-distributed reward structure.

	SARSA	EWA	dRL
Normal	**100**	0	0
Pareto	20	23	**57**

and hence is *reasonable* in responding to *typical* domain outcomes.As we show above in the certainty-equivalence task, our algorithm demonstrates a bias towards atypically large rewards and aversion to atypical losses characteristic of human decisions [32]. While this costs it some performance in normally distributed environments, it also affords it the opportunity to gain large rewards and/or avoid tail-risk when rewards are distributed according to heavy-tailed distributions. To demonstrate, we evaluated performance in the grid world with rewards drawn from Pareto distributions. Here cognitively efficient learning's bias towards extreme events causes it to perform better than its competitors (see Fig 5.3(b)). Further, the difference in cumulative predicted payoff is not a consequence of cognitively efficient agents obtaining very large rewards in a few decision instances. Table 5.1 shows that our algorithm outperforms its competitors in a large majority of individual trials, while being dominated by SARSA for normally-distributed reward samples under the exact same statistical testing conditions. Since many natural probability distributions are non-stationary, we believe that our results show how cognitively efficient learning is naturally robust under environmental statistics typical of real environments[2] .

5.5 Discussion

To see how the approach we have presented fits into the existing reinforcement learning literature, considere Table 5.2 where we categorize learning algorithms according to their reward assumptions: intrinsic vs. extrinsic and relative vs. absolute.

In recent years, a lively discussion has sprung up in the RL community over the possibility of introducing biologically natural forms of rewards, e.g. novelty, salience

[2] A deeper analysis of the reasons for this robustness is a current topic of research for us, but lies outside the scope of this thesis.

	Absolute rewards	**Relative rewards**
Extrinsic reward	standard RL [8]	Non-symbolic action representations [128]
Intrinsic reward	Intrinsic motivation [111, 121]	Motivated learning [16]

Table 5.2: Exemplars of reinforcement learning algorithms following different *effect-goal* assumptions

etc. into the objective function for a learning agent. Fundamentally, almost all such approaches attempt to model intrinsic motivation as secondary additive rewards which are obtained by computing some notion of *value* of the policies followed by the agent. Hence, a reward function of the form $r(x, a) = q(x, a) + D(u, p^*)$, where u is the agent's action policy and p^* the Platonic optimal policy, is justified to explain these approaches. The measure D reflects the difference in predictive performance through following two different policies. Singh and Barto's basic approach, which has lately been further developed by [121], assumes a level of intrinsic motivation proportional to prediction error that occurs through following the learned policy u as opposed to the optimal policy p^* [111], in consonance with our characterization of environmental cost. Simsek [118] adopts a more sophisticated approach by assuming that the value of the optimal policy is not known *a priori*, implying that the best use of the difference measure is to attempt to incrementally improve the value of the current policy. Towards this end, she defines the policy value of a policy $V(u) = \sum_{x \in \mathcal{X}} \rho(x) V^u(x)$, where $\rho(x)$ is the original state distribution and $V^u(x)$ is the state specific value. She further uses the difference between policy values to define D as

$$D(u_1, u_2) = V(u_1) - V(u_2).$$

Simsek uses D as the value function for a second MDP that is solved in parallel with the original MDP to solve the optimal exploration problem in a sequential decision-making setting, viz. if immediate reward-seeking is not a goal, what is the most efficient way of learning the domain's dynamics to maximize future reward.

Gershman's recent work [6] presents an interesting alternative statement of the intrinsic motivation problem. He assumes that there are neural computational costs associated with all decisions, and that an agent's intrinsic motivation is to find a low-cost approximately optimal policy that maximizes reward. He assumes the existence of given state-action specific reward functions $r(x'|x,a)$, and maximizes a variational lower bound on the log expected reward. This function has the form

$$E_u[\log r(x,a)] + E_u[\log p(x)] - E_u[\log q(x)],$$

which can be rewritten as

$$E_u[\log r(x,a)] - KL(u,p).$$

However, Gershman also includes an abstract multiplicative computational cost $C(u)$ into the reward term, causing the final objective function to look like

$$E_u[\log r(x,a)] - KL(u,p) - \log C(u).$$

The first term reflects the state-associated reward/cost-to-go which is maximized/minimized depending on interpretation, the second term reflects the environmental costs of adopting a preference that does not match the natural dynamics of the environment and the final term reflects costs entirely associated with properties of the policy representation itself. While this approach uses a decidedly sophisticated model of internal cost computations, it is based on the assumption of absolute and external reward.

Srivastava [16] have proposed a self-motivated learning objective of the form

$$T + \frac{\sum R(u,p_m)}{n - H(u)},$$

for a learning algorithm that shares our epistemic assumptions about reward. The approach described in our paper significantly deepens this work by giving an optimal model-based learning approach based on the same assumptions of intrinsic, relative reward.

Friston's free energy method [5] is the only other existing approach to our knowledge that uses the idea that rewards are entirely intrinsic. The free energy approach assumes the existence of a set of perceptual states $\mathcal{S}$, which the agent has an evolutionary/empirical probability λ of visiting. Agents generate a recognition density, another

probability distribution γ on the same states. Both λ and γ depend both on the environmental states encountered s and some internal agential parameters ν. Friston minimizes a variational bound on $-\log \lambda(s)$ of the form,

$$D(\gamma(\nu)\|\lambda(\nu|s)) - \log \lambda(s). \tag{5.15}$$

Friston's notion of reward deviates from ours, in that it is based entirely on the agent's encoding of the *statistical* properties of the environment. Thus, as [6] points out, a poor Fristonian agent will reject a winning lottery ticket because it is a statistically improbable outcome. Our agent, on the other hand, would still consider winning the lottery useful, if she *expects* to win it. Furthermore, our model's predictions on the certainty-equivalence task suggest that agents will overweight the probability of winning lotteries, precisely because such events are exceptionally rare, thereby being more inclined to expect winning than is statistically likely. Hence, the additional information-theoretic grounding of intrinsic reward sets our approach apart from Friston's free energy methods, while sharing in many of its neuro-cognitive motivations [5].

5.6 Conclusion

In this paper, we have built upon the theoretical structure of reinforcement learning to construct a more realistic model of goal-directed learning. The principal novelty of our approach lies in its understanding of relative preference for particular outcomes emerging intrinsically from the cognitive decision-making process of the agent. By assuming cognitive efficiency as a natural mechanism for memory recall, we instantiate a tractable learning model and retrieve preliminary experimental results in accordance with observed behavior in human subjects. We expect the results of our approach to contribute both in the future development of more realistic autonomous agents and in studies of human memory recall and choice selection.

Chapter 6

Cognitive efficiency as a causal mechanism for social preferences

6.1 Introduction

Sociology and economics both purport to study the processes by which people make decisions in the real world. However, their objects of study are conceptually exclusive. While sociology imputes proclivities to groups of agents, neoclassical economists assume that the decisions of groups emerge from the activity of self-interested individual actors. In practical terms, while much of modern economic theory is built around the *homo economicus* understanding of rational self-interest [87], computational social science modelers currently use heuristics built around the notion of social utility [129] like reciprocity [130], fairness [131] and social image protection [132] to establish what they consider to be more realistic objectives for social agents to optimize. While such heuristics are predictively useful, they provide limited causal insight into the individual level decision processes that presumably equilibrate to create known patterns of social behavior. This state of affairs makes it very difficult for economists to use incontrovertible sociological insights in their models, since the causal mechanism by which the interaction of self-interested agents creates patterns of social behavior remains unknown.

While it is unrealistic to expect the existence of rigidly mechanistic explanations of behavior in biologically realistic agents, given that humans share a common biological mechanism for assessing the value of social interactions, it appears natural to

hypothesize that the neurocognitive mechanisms of value appraisal, memory recall and choice selection play an important role in the generation of social preferences. As such, we consider it likely that a theory of decision-making that is grounded in a functional understanding of these neurobiological activities would allow us to better understand both the agency of social agents and the structure of social behavior patterns. Even more importantly, we expect such an exercise to give us a realistic generative model of social preferences delineating the precise nature of the interaction between *structure* and *agency* in social behavior. That is the task we undertake in this paper.

Srivastava and Schrater [16] have recently provided a principled causal account of human behavior in sequential decision-making tasks that presents a joint account of multiple families of cognitive 'biases' observed in human subjects. Our approach is premised on the assumption that intelligent organisms have evolved to make choices that involve minimal cognitive processing, while satisfying their needs. In this paper, we show that an extension of our cognitively grounded model of decision-making predicts well-known effects documented in the literature on social preferences. In particular, we show that cognitively efficient agents naturally prefer fair outcomes in ultimatum games, prefer to cooperate in prisoners' dilemma settings, prefer to associate with other agents that (a) they agree with or (b) resemble, and prefer to restrict their communications to within small groups.

6.2 A cognitive principle of least action

Consider a biological organism that is capable of observing its own preferences with respect to the environment[1] , but which needs access to resources in the environment in order to retain energetic homeostasis. Assuming that resource availability fluctuates in both space and time, satisfactory communication with the environment effectively becomes a prediction task, with the organism's goal being constructing theories of the environment sufficiently predictive to secure enough resources to ensure survival of the genotype. Furthermore, over generations, we expect selection pressure to promote efficiency in the use of limited cognitive resources in a population of such agents. In light of

[1] In other words, a metacognitive [108], or self-aware organism

this understanding of metacognitive intelligence as being a fundamentally predictive organ, we suggest that minimizing cognitive effort in constructing sufficiently informative beliefs about the environment is a general principle of intelligent behavior for humans in particular, and all metacognitive organisms in general.

While on the one hand, this alternative account of what constitutes intelligent behavior is evolutionarily plausible, it has the further advantage of requiring weaker axiomatic assumptions than classical economic theories of belief formation. Given ordinal preferences, such theories assume the validity of the von Neumann expected utility hypothesis [34] to obtain cardinal reward values. Basic rational choice theory assumes that rational agents attempt to maximize the reward that they can obtain through their actions. This canonical framework is formalized in learning theory as reinforcement learning [8]. However, the expected utility hypothesis has been shown to be unrealistic in multiple behavioral studies (see [84] for a general review) and is now widely acknowledged as being deficient. In our framework, environmental phenomena are judged to be valuable to the extent they have been judged valuable in the past. Judging utility by *whether* an option has been useful in the past as opposed to *how* useful it is removes the necessity to postulate cardinal rewards embedded in the environment. Hence, as we show below, the cognitive efficiency hypothesis leads to a principled and realistic model of decision-making that is independent of the von Neumann-Morgenstern assumptions.

We provide a brief description of our model of the choice selection of individual agents below in 6.2.1. The key idea in our approach is to formulate the generation of future expectations from a cognitive model of memory in lieu of statistical expectations, leading to interesting and realistic choice predictions. The goal of agents operating in our framework is to minimize cognitive processing costs while retaining predictive confidence about their environment. Then, in Section 6.2.2, we extend this model of agency to account for social cognitive decision-making, and hence obtain a positive model of social preference formation.

6.2.1 A cognitively efficient decision model

We now briefly describe our decision-making model[2] . The core premise of our approach can be formalized in the following way: an agent tries to minimize its cognitive

[2] A fuller description is available in [7] and [16].

processing cost T while maintaining a 'satisficingly' high level of predictive confidence C in the quality of its choices. The cognitively efficient learning objective is then seen to be identical with minimizing a function of the form,

$$\begin{aligned} \underset{\mathbf{x}}{\operatorname{argmin}} \quad & T \\ & C_{\text{new}} \geq C_{\text{old}}. \end{aligned} \tag{6.1}$$

where T and C are quantified below in terms of beliefs.

Let the discrete probability distribution $\mathbf{x}(s)$ represent an agent's belief about the relative quality of outcomes $s \in \mathcal{S}$ available to it. As the agent interacts with its environment, its belief changes in a way that allows it to maintain biological homeostasis. However, since the agent must necessarily learn the dynamics of the environment over and over, it must continually experience regret. Unlike the traditional definition of regret used in the online machine learning literature [96] which compares obtained results with a Platonic ideal result, we define regret as a quantity centered around the agent's current belief. Mathematically, a measure of regret experienced by an agent operating with a belief $\mathbf{x}_a$ for a different belief $\mathbf{x}_b$ can be quantified with an information divergence [69] of the form,

$$R(\mathbf{x}_a, \mathbf{x}_b) = \sum_{j=1}^{n_a} \mathbf{x}_a^j(s) \log \frac{\mathbf{x}_a^j(s)}{\mathbf{x}_b^j(s)}. \tag{6.2}$$

The information divergence measure is intuitively suitable for representing differences between beliefs encoded as probabilities, since it is asymmetric and non-metric. The asymmetry leads to the current belief being privileged in a particular intuitively sensible way (easy to find past belief that is closest to current belief, converse is hard). The non-metric nature of the information divergence allows for intransitive selection between gambles, as seen in the Ellsberg paradox, for instance, to occur.

We model cognitive processing costs as the cost of recalling past beliefs into memory. Unlike existing computational techniques, which multiplicatively discount past experience [8], we argue that natural agents find exceptional beliefs easier, and hence less costly, to recall. An agent trying to predict efficiently can measure exceptionality as deviation from the average level of regret it experiences in making its predictions. Notably, recent neurobiological experiments [126] provide some empirical support for

the particular definition of exceptionality we hypothesize. Hence, we measure the informational exceptionality of a past belief x_{old} (and hence the ease with which it will be available for recall to the agent) as the deviation from the average surprise experienced by the agent R':

$$A(\mathbf{x}_{\text{old}}) = |R(\mathbf{x}, \mathbf{x}_{\text{old}}) - \overline{R}|, \tag{6.3}$$

where x is the agent's current belief.

Given ease of memory access for each past belief, a reasonable measure of the processing cost of selecting a subset $\mathcal{M}'$ out of the set $\mathcal{M}$ of all past beliefs is the inverse exceptionality-weighted sum of the nominal cost of accessing all beliefs in $\mathcal{M}'$. Assuming the nominal cost of accessing each belief to be unity, the total cost of memory access T becomes,

$$T = \sum_{\mathbf{x}_i \in \mathcal{M}'} A^{-1}(\mathbf{x}_i), \tag{6.4}$$

Our measure of the agent's confidence in its ability to predict its environment, $C : x \to [0, 1]$ captures the idea that confidence grows when the beliefs have low uncertainty and low surprise:

$$C = \frac{1}{C_{\max}} \frac{\log |\mathbf{x}| - H(\mathbf{x})}{\sum_{\mathcal{M}'} R(\mathbf{x}, \mathbf{x}_{\text{old}})}, \tag{6.5}$$

where the numerator is a monotonically decreasing function of the Shannon entropy $H(x)$ of the belief. Note that C is normalized with respect to the greatest value it has previously been observed to achieve.

Any algorithmic solution of our model's objective function must solve three problems. One, we must specify a memory update specifying how prior beliefs are combined to produce the agent's current belief. We formulated this as a race-to-threshold model [cite], where old beliefs populate active memory with a latency proportional to their exceptionality. This results in a memory update of the form,

$$m_i(s) = \frac{1}{\max(1, |\mathcal{M}'|)} \sum_{k=1}^{|\mathcal{M}'|} \mathbf{P}_k \; x_k(s) + \mathbf{N}_k \; \overline{x_k}(s), \tag{6.6}$$

where $\mathbf{P}$ is an indicator vector that takes values 1 for low regret salient instances (zero otherwise) and $\mathbf{N}$ takes value 1 for high regret salient instances (zero otherwise). The notation $\overline{m}$ represents inverting the set of beliefs under consideration and can be obtained in several ways, e.g. subtracting each component value from 1 and renormalizing.

In cases where the salient set is empty, m_i simply takes on the value of the immediately prior belief x_{i-1}, reflecting the intuition that no memory recall took place.

Two, we must specify an environmental update, which shows how the agent obtains information about the environment and integrates it into its current belief. For the case of individual decision-making, we assume that sensory data is encoded into the space of possible outcomes as a relative preference by evolutionarily adapted perceptual processes.Our usage of the term *reward-inference* accentuates the fact that this information is obtained after perceptual processing of environmental stimuli. In [7], we show that a statistically optimal and cognitively plausible mechanism for updating our agent's quality-belief would involve a convex sum between the existing quality-belief and the incoming reward-inference signal. Since the existing quality-belief at every decision instance is identical with memory, at the i^{th} decision instance, we take the current agent quality-belief to be calculated as,

$$x_i(s) = C_i m_{i-1}(s) + (1 - C_i) g_i(s). \tag{6.7}$$

Three, we must specify a combinatorial optimization algorithm specifying which subset $\mathcal{M}' \subset \mathcal{M}$ of existing beliefs in memory the agent will *recall* to form its new belief, such that the objective function we have defined above is optimized. It is quite straightforward to find an optimal $\mathcal{M}'$ while respecting the confidence constraint by incrementally populating $\mathcal{M}'$ from a list of beliefs sorted by exceptionality. The set $\mathcal{M}'$ is then used to construct the memory update (6.6). The memory update, along with current values of reward-inference and confidence, is used to construct the optimal quality-belief for the relevant decision instance using (6.7). We thereby obtain a learning algorithm for predicting choices made by agents in sequential settings. The resultant algorithm outputs beliefs corresponding to an agent's relative preference for each of the possible outcomes in a particular decision context.

6.2.2 A model of social preference learning

The decision-making model presented in SI explains individual decision-making by obtaining information about the environment through the organism's perceptual apparatus

in the form of reward-inference. However, in a purely social[3] environment, such information can be assumed as being provided by other agents. In this case, we would interpret $g(s)$ as encoding *guidance* provided by other agents through their own revealed preferences over a shared set of possible outcomes.

Since the agent can now use the beliefs of other agents to form its decisions, it can continue to be decisive even if its own predictive confidence with respect to the decision context is low by *trusting* other agents. Therefore, we generalize predictice confidence to encompass the confidence an agent will have about the predictive value of the guidance provided by another agent.

Assuming a particular social environment populated by $a \in \mathcal{A}$ agents, the confidence of agent a in agent a' is defined in a manner analogous with the agent's individual predictive confidence definition, viz.,

$$C^{(a,a')} = \frac{1}{C_{\max}} \frac{\log |g^{(a')}| - H(g^{(a')})}{\sum_{\mathcal{M}'} R(g^{(a')}, x^{(a)}_{\text{old}})}, \tag{6.8}$$

In the social preference learning model, agent a will update its belief as,

$$x_i^{(a)}(s) = C_i^{(a)} m_i^{(a)}(s) + \sum_{a' \in \mathcal{A}_i, a' \neq a} C_i^{(a,a')} g_i^{(a')}(s), \tag{6.9}$$

where x is subsequently normalized and $\mathcal{A}' \subset \mathcal{A}$ is the subset of agents that provide guidance to agent a at a particular decision instance. Finally, communicating with other agents to obtain their revealed preferences involves some cognitive processing, which must be incorporated in the agent's overall cost function. Let the communicative cognitive cost incurred by agent a in obtaining information from agent a' be $T_c^{(a,a')}$. Extending our earlier arguments from Section 6.2.1, it follows that a rational agent in a social setting self-selects both $\mathcal{M}'$ to reduce costs of memory recall and $\mathcal{A}'$ to reduce cognitive costs of communication while maintaining a sufficiently high predictive confidence.

Hence, the optimal decision problem in social settings can be represented as a dual

[3] By purely social, we mean an environment with weak perceptual bases for forming decisions, e.g. populist discretionary consumption choices.

subset selection problem of the form,

$$\begin{aligned} \underset{\mathcal{M}',\mathcal{A}'}{\operatorname{argmin}} \quad & \sum_{\mathbf{x}\in\mathcal{M}'} A^{-1}(\mathbf{x}) + \sum_{a'\in\mathcal{A}'} T_c^{(a,a')} \\ C^*_{\text{new}} \quad & \geq C^*_{\text{old}}, \end{aligned} \tag{6.10}$$

where, $C^* = \max(C^{(a)}, \max C^{(a,a')}), a' \in \mathcal{A}'$.

Note that this model of social preference learning retains the cognitive principles of the individual decision-making model we have described above. The only change is in the nature of the reward-inference/guidance term. Whereas the model described in [7] imbues this term with a perceptual interpretation internal to the organism, in the social model, it is equated with the revealed preferences of other agents obtained via communication. In a purely social domain, an agent will obtain guidance purely through social communication. However, it is straightforward to generalize (6.9) to a scenario where multiple guidance terms reflect reward inference obtained from multiple perceptual modalities within the same biological agent, while still others represent the revealed preferences of other agents.

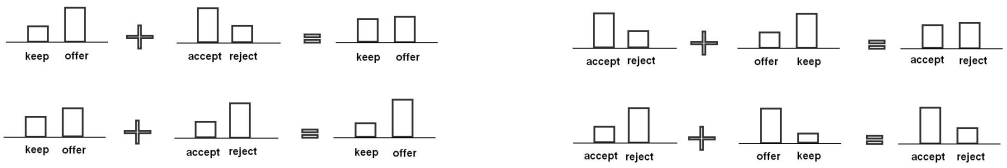
6.3 Experiments

The experiments reported in this paper represent a first attempt at demonstrating a general theory of social preference formation. As a consequence, we have used the simplest and most abstract decision domains feasible. While such a selection may be criticized for its lack of sophistication and realism, we suggest that it is far more important for inchoate frameworks such as ours to demonstrate their value in as clear and interpretable terms as possible. We have therefore restricted our experimental settings largely to game-theoretic settings historically favored by experimental economists in deriving alternative theories to *homo economicus* rationality. In particular, we show how agents using our cognitive efficiency criterion make choices that better resemble the choices of human subjects in the ultimatum game, in the iterated prisoners' dilemma game and in generic social link formation.

6.3.1 Inequity aversion in ultimatum games

In the ultimatum game, sequential decisions by two players determines how they will split a fixed pot of money M. The first player chooses some amount, say p for himself, a partition which his opponent can choose to either accept or reject. If the second player accepts the split, he gets $M - p$, and the first player gets p. If he rejects the split, both players get 0. It is easily shown that the optimal strategy for the first player (proposer) in such games is to estimate, over repeated trials, the maximum p one can keep for oneself that will not provoke a rejection from the other player. At the same time, the other player (selector) would do best to reject substantially low offers, since not doing so would drive the bid even lower.

The ultimatum game is of interest because, whereas the expected rational strategy for the first player (proposer) is to make the lowest feasible offer to the other player, human subjects across cultures tend to make offers close to $p = M/2$, thereby displaying inequity aversion [131]. Furthermore, whereas the rational strategy for the second player (selector) would be to accept any non-zero offer, human subjects frequently reject unequal offers in violation of the *homo economicus* sense of rational self-interest.



(a) Belief updates for proposer

(b) Belief updates for selector

Figure 6.1: Belief updates for self-motivated social agents in an ultimatum game. Note that the state ordering is switched between the two vantage points, reflecting the mirrored motivational structure. We disregard the influence of the confidence term in these diagrams. For actual updates, confidence in the antagonist will proportionately affect the degree to which an agent's decision will be affected by new guidance provided.

Normalizing the total payoff, we can transpose the space of payoff splits into a binary probability space, so that the proposer's belief can be represented as binary probability $\{p, 1 - p\}, p \in [0, 1]$. The selector's decision is to either accept or reject the proposal.

For formal tractability in our framework, we assume that the latter decision is also represented by a binary probability $\{q, 1-q\}, q \in [0,1]$, with proposals rejected for $q < 0.5$ and accepted otherwise. We further define the minimum payoff threshold p_m such that $q(p_m) = 0.5$. A learning agent interacting with other agents via the ultimatum game is expected to learn realistic values of p_m and/or the function $q(p)$, depending on their role in the interaction.

The ultimatum game integrates into our framework as follows: the proposer communicates p to the selector; the selector communicates q to the proposer. Both forms of communication reflect agents' beliefs about the same set of outcomes and hence additively update the respective agent's belief using (6.9), with the general structure of the update following the schemata shown in Figure 6.1.

In simulation, in the case of the proposer, low regret instances arise when the proposal is accepted; high regret instances arise when it is rejected. Thus, according to our model, the proposer will seek out any payoff it can obtain while not accumulating regret. On the other hand, the selector will experience high regret when it rejects a proposal, and will experience low regret when it accepts. It will therefore, preferentially accept, unless the proposal is too unequal, which makes the high regret term deeply exceptional, causing it to be prominently activated in active memory and influencing the selector to reject the proposal. If we allow both players to learn both roles symmetrically, viz. with equal number of trials as proposer and selector, we find that the proposer's payoff expectation equilibrates to $p_m = x/2$.

Recently, [133] has showed that the canonical Fehr-Schmidt model [134] fails to account for differences in outcomes in ultimatum games that take the intentionality of the players into account, and has suggested a modified procedural Fehr-Schmidt model, where agents are sensitive to payoff expectations, as opposed to immediately observed payoffs. It is evident that prospective expectation is a better measure of social utility than absolute value, as it takes process intentionality into account. However, while [133] correctly recognize that agents' assessment of the process is a better reward estimate than immediate values obtained, they fallaciously assume 'expectation' to mean the same cognitively as it does probabilistically. As we discuss in detail in [16], there is reason to believe that humans use an alternative procedure to compute cognitive expectations, one that our model of decision-making captures. As a result, our model

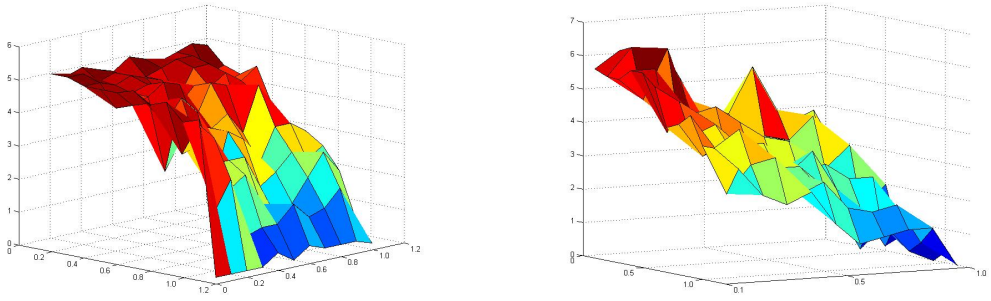
of social utility, while in functional accordance with the process view of FS utility, as demonstrated by our results in Section 6.3.1, differs in its mathematical formulation. Interestingly, our results provide negative evidence for the existence of any intrinsic value in 'fairness', as has been suggested in some earlier studies. An egalitarian environment, characterized by an equal number of trials as proposer and as selector for each agent, leads to equilibration around the symmetric split $p_m = x/2$, whereas a skewed training period leads to equilibration around asymmetric splits. This leads to the somewhat disheartening conclusion that humans do not, in fact, possess an intrinsic high valuation for 'fair' outcomes. Considering the enormous inequalities present in economic systems now and throughout history, this should not, however, be a surprising conclusion. Whether multiple equilibria can be characterized in the case of asymmetric splits and whether the degree of asymmetry in a relationship can be modeled as a function of their interaction history remain open questions for future research.

6.3.2 Reciprocal behavior in iterated prisoner's dilemma

In this experiment, we demonstrate the emergence of super-rational [135] behavior in our model's predictions in an iterated prisoner's dilemma (PD) setting.

The prisoner's dilemma problem has been extensively studied in the game theory and sociology literature [136]. In the basic PD setting, the utilitarian rational strategy for either player is to defect, which, sadly, leads to a poor outcome for both players. Blindly cooperating, however, is even worse, which means that a player must have a responsive strategy for dealing with an opponent. Playing multiple PD games with the same parameters affords the opportunity to discover the opponent's strategy and adjust one's own. This makes iterated PD an interesting problem domain for testing agents that purport to behave in a manner consonant with autonomously motivated humans.

It is possible to describe the space of strategies in two player PD games as a set $\mathcal{S}$ of tuples $S(p1, p2)$, where $p1$ is the probability with which the agent defects if the opponent cooperated and $p2$ the probability that the agent defects if the opponent defected on the previous turn. We evaluate the performance of two different agents: one using a hard-coded tit-for-tat (TFT) strategy (cooperate if the other player cooperated last turn, defect otherwise), the other beginning agnostically and learning an appropriate strategy for dealing with the agents it encountered using self-motivated learning. We



(a) Payoff surface for a tit-for-tat strategy

(b) Mean payoff surface for a strategy learned by a self-motivated agent over 50 trials

Figure 6.2: Results for prisoners' dilemma experiment

assessed the performance of both agents against a grid of mixed strategies obtained by varying $p1$ and $p2$ between 0 and 1 in increments of 0.1 along two axes.

Figure 6.4(a) shows the payoff obtained by TFT. To briefly orient the reader, the rightmost corner represents the case where the opponent's strategy is to always defect, i.e. $S(1,1)$ which causes both players to continually defect (tit-for-tat) and obtain low payoff. On the other extreme, the *saintly* strategy $S(0,0)$ lives in the leftmost corner. Here, both players continually cooperate and receive the intermediate payoff.

Figure 6.4(b) shows the average payoff obtained by our model over 50 iterations on each grid point. Our model learns, for the most part, a strategy that closely resembles TFT. The TFT strategy is well-known in the PD literature both for having been postulated as a model of human behavior as a theory of reciprocal altruism [137] and for being exceptionally robust as a game-theoretic strategy against other strategies in iterated PD games [136]. Unlike existing adaptive agents [138, 139], our agent does not possess an explicit model of its adversary's choices. Its preference to cooperate, therefore is intrinsic, not game-theoretically planned. The motivation to select the cooperative option instead of the defect option inspite of a lower extrinsic payoff is explained by the additional intrinsic payoff obtained by the agent not having to continually change its prediction (and incurring a cognitive cost) once it has begun cooperating.

We note, in passing that our model's learned behavior deviates from TFT in interesting ways. For example, in cases where the opponent is too saintly and does not

retaliate, our strategy learns to exploit it by electing to defect continually. Further analysis of these and other deviations of our model's predictions from TFT presents an interesting direction for future work.

Our experiments show that cooperation emerges as a natural response strategy for cognitively realistic decision-makers. This finding suggests cognitive efficiency as a causal mechanism for learned altruistic behavior, i.e., the cost of predicting other agents' behavior rises under antagonistic choices, causing altruistic choices to be preferred. This simple explanation presents a parsimonious mechanism for the development of learned altruism [137]. Reciprocal altruism, as seen in tit-for-tat repeated PD games, is widely acknowledged as a powerful ultimate explanation for human altruism in small and stable groups [130]. One of the principal criticisms levied against direct reciprocity theories is that they have heretofore assumed that agents cooperate in anticipation of future reward. Such a utilitarian mechanism cannot explain the emergence of strong reciprocity - cooperative actions performed in the absence of external reward - both in human and animal subjects [**?**]. Our experimental results demonstrate that it is possible for completely intrinsic payoffs to promote the development of approximately reciprocal strategies in two-player games, thus potentially resolving the utilitarian critique of strong reciprocity. This view is further supported by evidence from neurobiology that suggests that individuals experience particular subjective rewards from mutual cooperation [140].

6.3.3 Homophily, groupthink and preferential attachment in social link formation

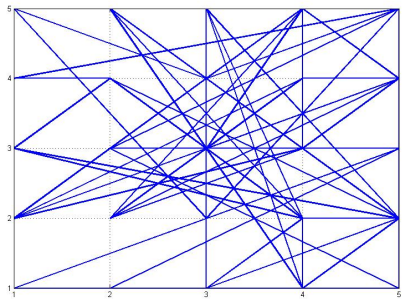
In sociology, homophily is the tendency of individuals to associate and bond with similar others. The presence of homophily has been asserted in a large number of studies (see [141] for a comprehensive review). While multiple culturally specific hypotheses about the cognitive processes underlying homophily have been documented, we suggest they can all be subsumed within the general hypothesis that (i) shared social knowledge or features facilitates communication, (ii) ease of communication determines attraction implying that (iii) social similarities cause attraction. This hypothesis is well-supported in the existing literature. For example, *constructuralism*, a prominent sociological theory outlined by [142] fundamentally assumes that people who share knowledge with one

another are more likely to interact and form ties. Demographic and cultural [141] similarities have been repeatedly shown to be correlates of shared knowledge and hence for ease of communication, thereby potentially causing homophily.

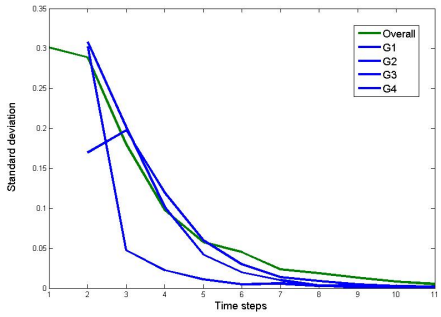
Our cognitive cost minimization framework supports the communication facility hypothesis for the emergence of homophily. Agents attempting to minimize their cognitive costs while selecting amongst multiple agents to interact with are likely to select agents whom they can communicate with more easily, thereby reducing cognitive costs. To test this hypothesis, we set up a simulation containing agents $a \in \mathcal{A}$ each associated with a non-unique D-digit binary code[4] . Following the intuition that socio-cultural similarities are correlated with shared knowledge and ease of communication, we define the communication cost between agents a_1 and a_2 as $T_{1,2} = \sum_{i=1}^{D} m_1^i \oplus m_2^i$. At every step of the simulation, agent a can choose to communicate with any other agent $a' \in \mathcal{A}$, but will incur a cost $T_{a,a'}$ in doing so. Each agent maintains a random belief $x \sim Bern(p), p \sim U(0,1)$, which also becomes its guidance to other agents during communication. Figure 6.3 shows some results from this simulation. As expected, agents sought out low communication cost counterparts to preferentially link with, resulting in a sparsification of graph structure into clusters of well-connected components (see e.g. Figure 6.3(a)). Furthermore, we find that, in order to further reduce cognitive costs of repeated interactions with similar neighbors, group members equilibrate their preferences to a common value, as shown in Figure 6.3(b). This observation points to a possible causal connection between homophily and herding behavior or groupthink.

A testable prediction arising from this interpretation of homophily is that such homophilic behavior is likely to occur only in the incipient stages of social group formation, or the entry of a newcomer to an existing group. Once shared representations have been created, the marginal value of preferring one representation over another will be minimal, causing the marginal costs of belief divergence to dominate agents' decision-making. We therefore expect co-prediction in significant decision contexts to be a stronger predictor of social tie formation among mature groups and older individuals. Significantly, the first half of this hypothesis is borne out by studies showing association with similar others being more significant than actual socialization as an indicator of tie formation among adolescents [144].

[4] This code may be interpreted as a feature space representation for socio-cultural factors.

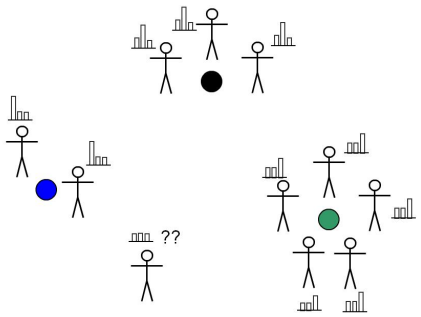


(a) Sample social graph generated from an originally completely connected population of N = 25 agents. Note hub-like structure and asymmetry in degree. Interestingly, a distinct central clique emerges in this particular simulation run, resulting from similar identity codes being assigned to a significant fraction of the population.

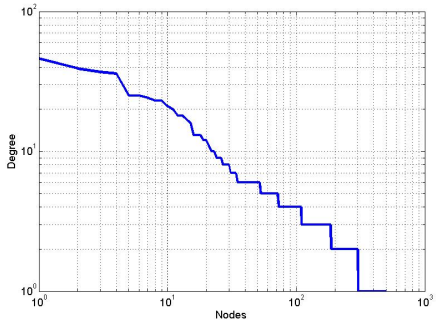


(b) Inter-group standard deviation is compared against population standard deviation at every time step for four major groups. The largest group (G4) shows quick convergence to a preference value, while the others converge slowly, but surely. Low inter-group standard deviations show that preferences converge to particular values within groups disjoint from the preferences of the larger population, thereby demonstrating groupthink.

Figure 6.3: Cognitively efficient agents seek to minimize communication costs by preferring linkages with individuals that resemble them. In this process, they end up holding preferences that closely resemble those of their similar peers, resulting in further cognitive cost reduction in future interactions. These results support Kosslyn's social prosthesis hypothesis [143] while concomitantly showing a causal link between homophily and groupthink.



(a) Schematic showing the social choice an agent faces when encountering an existing population with established belief systems.



(b) Degree distribution of 500 agent population grown incrementally from a 10 agent seed population using the social preference learning model. Tail of distribution fits a power law with a scale factor of 2.7

Figure 6.4: Agents using internal cognitive costs to assess rational choices attach to existing belief systems preferentially, replicating the generative process of Barabasi and Albert [145]

We further note that our simulated agents display homophily and preferential attachment behavior simultaneously. To see this more formally, consider an initial population of k agents, with each agent holding one of m possible beliefs, $m << k$. Agents that hold the same belief are homophilic and form hubs. The $(k+1)^{th}$ agent entering the network will now align his beliefs with any agent it encounters with uniform probability. However, since homophilic agents share beliefs, the probability of sharing the belief of a hub becomes proportional to its degree n_m. It is straightforward to show that this generative process immediately converges to a BA model of link formation [145] widely recognized in social networks.

6.4 Discussion

The debate concerning the primacy of either structure or agency with regard to human behavior characterizes a central ontological dispute between dominant paradigms in sociology and neo-classical economics. In this context,'agency' refers to the capacity of individuals to make choices based on their autonomous motivations. 'Structure', on

the other hand, refers to the existence of specific patterns of relationships that predispose individuals to behave in the ways they do. Neo-classical economics is built upon the *homo economicus* premise that actors are autonomously motivated self-interested agents. On the other hand, while sociology has a rich history of agent-based modeling, the most recent wave of computational sociology [146], rather than employing simulations, uses network analysis and statistical techniques to analyze large-scale computer databases of electronic proxies for behavioral data. Such efforts are naturally better aligned with structuralist theories of sociology, and have led to an increased emphasis in discovering patterns of group behavior at the expense of realism in modeling agential motivations.

An attempt to reconcile these two paradigms is observed in the theory of structuration [147], which holds that agents make decisions within the context of a pre-existing social structure which in turn, emerges adaptively through the choices of agents. [148] have shown that such a framework is well-described by a reinforcement learning strategy that allows an agent's behavior to be affected by their self-perception of how well they are making predictions. Our algorithm follows [148] insofar as our definition of self-perceived confidence also depends on prediction accuracy. Our approach differs from [148] in allowing both highly expected and unexpected experiences to produce emotional affect. Whereas earlier results in the literature on dopamine response to prediction errors [105] suggested that only unexpected results yield intrinsic motivation and hence emotional affect, more recent discoveries [126] suggest that the dyadic structure we have postulated is more plausible. We further note that our model replicates a central finding of [148] that agents 'feeling positive' can think ahead in a narrow sense and free up working memory resources, while agents 'feeling negative' must think ahead in a broad sense and maximize usage of working memory.

The emergence of effects seen both in experimental economics games and in sociological theories from a cognitively grounded model of individual decision-making strengthens the case for using learning algorithms to resolve the structure-agency impasse. However, most existing machine learning algorithms make unrealistic assumptions about the nature of agency, which renders them unsuitable as models of social agents. By modeling the agency of biological organisms better, our approach provides a useful theoretical beginning towards understanding how agents and social preference patterns affect each

other. Testing the predictions of our models on real-world longitudinal social network data, e.g., from gaming environments, presents a fascinating direction for future work.

As [149] point out, although computational social agents are often characterized as being 'cognitive', there have been relatively few attempts to carefully emulate human cognition. Models of agents have frequently been specialized to problem domains, resulting in the use of highly domain-specific heuristics in lieu of actual autonomous decision-making. Sun and colleagues have been influential [cite Cog Sci 2010 workshop] in pointing out the necessity of grounding social sciences in theories of cognition. While heuristic-based models may suffice as predictive tools for various social dynamics applications, they do not provide much insight into the causal mechanisms that underpin the emergence of social preferences. For the latter purpose, computational models grounded in realistic models of individual cognition are essential.

However, existing efforts in creating this synthesis have been attached to a particular comprehensive cognitive architecture - CLARION [66] which uses a dual-layer representation of explicit and implicit knowledge to perform general cognitive tasks. Partially as a consequence of its connectionist origins, and partially because of its immense scope, interpreting the social simulations of CLARION is problematic. Thus, while the need for grounding social science in theories of cognition appears clear, existing approaches do not address it satisfactorily. Our work shows that a more parsimonious *rational* [42] model of cognition can generate realistic social preferences without having to resort to complicated cognitive architectures.

6.5 Conclusion

We have demonstrated that several well-known sociological patterns of behavior can be seen to emerge from the individually rational choices of metacognitive agents attempting to make cognitively efficient decisions. In particular, we observe that previously ungrounded definitions of social utility and multiple accounts of putatively irrational social preferences can be jointly explained as being the functional result of cognitive cost minimization strategies used by individual agents. Our results unify multiple hypotheses from experimental economics and mathematical sociology and provide an original

and neuroscientifically principled theoretical connection between current theories of individual and social preferences supplemented with multiple testable hypotheses.

Chapter 7

Conclusion and Future Directions

7.1 Summary of contributions

The principal contribution of this work is the quantitative elucidation of a natural law that purports to explain how humans' preferences for different possibilities in the world change. The experiments and simulations described in this dissertation were performed to demonstrate the possibility of successfully abandoning epistemological fallacies that have confounded progress in quantifying behavior for the last 200 years, leaving behind understanding of a simple dynamic process of belief dynamics governed by a principle of least cognitive effort.

Since it is largely theoretical economists who have concerned themselves with modeling preference behavior quantitatively, and since expected utility maximization is the only competing rational explanatory principle for human behavior, the major portion of this dissertation is devoted to explaining economic data that utility maximization theories cannot explain. While efforts in this direction constitute an active area of research, no competing models succeed in devising *rational* explanations for the breadth of behaviors that our simple theory can explain. In particular,

1. We present the first *rational* explanation for the emergence of context effects in risk-free choices in Chapter 2.

2. We present the first *rational* explanation for the dynamic occurrence of prospect theory probability distortions in Chapter 3.

3. We present a unified *rational* explanation for three major families of cognitive biases - confirmation biases, primacy/recency effects and probability distortion biases in Chapter 4.

4. We present simulations demonstrating the equivalence of decisions from experience and decisions from description, reconciling two different paradigms of cognitive modeling of human behavior in Chapter 4. The existing literature treats both these regimes as qualitatively different. We show that they are not, and that both regimes can be *rationally* elicited from agents following cognitively efficient belief dynamics.

5. We present a generalization of reinforcement learning that is sensitive to cognitive architecture effects, displays robustness to non-stationary reward regimes and reproduces human behavior in interesting task settings in Chapter 5.

6. We show how individual cognitive efficiency leads to cooperative and altruistic behavior without explicitly forcing inequity aversion via social utilities on agents in game-theoretic environments in Chapter 6. These results are the first evidence for *strictly individual rationality* leading to altruistic behavior.

We emphasize further the significance of **rationality** in all the results in this dissertation. Not only does cognitive efficiency explain data across tasks, models and theories, it does so in a causally meaningful way. Rather than try to fit parameters in statistical models, as is standard practice in non-expected utility decision theories, we have confronted the problem of predictive behavior head on, and developed a normative theoretical framework *ex nihilo*.

Our adoption of this clear methodology renders the possible range of merits of our contribution transparent. If further research shows that this explanatory principle is insufficient in explaining belief dynamics, our general contribution is the quantification of cognitive effort in a meaningful way, leading to interesting economic predictions. If, on the other hand, it turns out that this hypothesis is sufficient to explain belief dynamics in human behavior, our general contribution is a redefinition of rationality to replace *homo economicus* as the standard for assessing human behavior across all behavioral disciplines.

7.2 Future directions

The ideas developed in this thesis, should they be proved correct, will leave scarcely any genre of behavioral research untouched. However, in the interests of concreteness, I have given below a list of possibilities that are personally interesting to me as future research problems building upon my dissertation research.

7.2.1 Experimental validation

The key assumptions underpinning the principle of least cognitive effort have received indirect support from other empirical studies. However, it is necessary to devise a specific set of behavioral studies specifically designed to falsifiably test them. In particular, it is necessary to verify whether:

- Humans' preferences in multi-choice tasks do, in fact, vary in the manner predicted by our value inference model in Chapter 2.
- Evidence accumulation in multi-choice tasks does, in fact, use context-sensitive desirability pointers, instead of standard utility measures.
- Both extremely typical and exceptional experiences are, in fact, easier to recall, as predicted by the cognitively efficient memory model in Chapter 3.

Verifying that each of these three behaviors arise robustly in human subjects will go a long way in affirming the correctness of our overall theory. Empirical testing of this principle is a necessary pre-condition for its suitability in the applications that we describe further below. In addition, such data will help us in further strengthening the basic individual decision theory and stimulate development of the as yet inchoate social decision theory outlined in Chapter 6.

7.2.2 Microeconomic applications

Microeconomic models of human behavior are fundamentally grounded in the well-known principle of supply and demand, viz. that in perfectly competitive markets, increased supply or decreased demand reduces prices, and vice versa. Why should this be the case? Clearly, in many practical cases, this is an intuitive observation. The price

of mangoes drops when they are in season and in greater supply, and drops during the lean season, as the supply-demand curve predicts. But why does this happen? Why do mango sellers reduce prices during times of plenty? The theoretical explanation for this occurrence proposes that market participants attempt to drive the market to a price equilibrium, an elegant explanation that, however, makes epistemologically challenging assumptions about their abilities.

Microeconomics research in the past 30 years has focused on identifying epistemic fallacies, e.g. information asymmetries [150], cartelization, etc. and their effects on market behavior, an exercise that has proved advantageous in explaining many deviations from optimal market behavior, non-zero transaction costs, etc. Further, the insights developed in such research has been used to inform more detailed and realistic economics models, whether of the dynamic stochastic general equillibrium variety, or agent-based systems. Ultimately, however, the logical culmination of the microeconomics enterprise would be the rapproachment of economic microfoundations with scientifically verifiable cognitive theories. Since this work represents one of the first such theories, it is clearly desirable to attempt to reconcile its predictions with the normative expectations of microeconomic agents.

To take a few concrete instances, our simulation results in Chapter 3 showed some preliminary predictions about individual differences in risk judgments, in line with empirical data. Further research along such lines is expected to identify conditions under which subjects behave as if following expected utility theory, cumulative prospect theory and/or other heuristics, as described in recent studies, as well as yield predictive models of individual differences in risk sensitivity based on task experience. Similarly, the results in the multiple price list simulations in Chapter 2 are expected to yield predictions about individual differences in risk aversion traditionally explained using wealth effects. A successful model of this process would reconcile an extremely problematic theoretical and empirical divide between expected utility theories of income and expected utility theories of wealth. Finally, and perhaps most intriguingly, our development of history-sensitive value-inference yields a natural representation both for Knightian or radical uncertainty and for Rothbardian utility-free analysis of behavior. Thus, our theoretical framework provides the mathematical tools for an unprecedented quantification of heretofore qualitative Austrian economic insights.

7.2.3 Macroeconomic applications

In a desire to quickly address Lucas influential critique [151] , neoclassical economists in the 1970s summarily abandoned the adaptive-expectations view of macroeconomics in favor of structural modeling using rational expectations. Thus, in place of assuming that variables of economic interest are learned as sample means of past values, neoclassical economics began assuming that expected variable values were always model-optimal. As is plainly evident with the benefit of hindsight, the mathematical legerdemain of the resulting rational choice theory elides the fact that the assumption of rational expectations has weak statistical basis. Market prices clearly suffer from systematic distortions that cannot be explained either in the rational or the traditional adaptive expectations framework. Interestingly, Lucas own suggestion for resolving his critique was to try to model deep parameters like preferences, resource constraints, viz. developing realistic microeconomic models by modeling psychological variables that describe the behavior of individual agents instead of trying to model abstract economic variables that describe how the entire economy works . In short, his point was that the adaptive expectations approach was not wrong; it was just modeling the wrong variables. In eschewing adaptive expectations entirely in favor of an unrealistic axiomatic basis for macroeconomic modeling (see e.g. [152]for a prescient technical critique, [153] for an excellent non-technical exposition), neoclassical economists appear to have thrown the baby out with the bathwater, resulting in an ever-widening gulf between economics theory and real-world activity.

In recognition of this state of affairs, while significant advances have been made in the development of modeling techniques designed to accommodate stylized facts concerning the putatively irrational behavior of financial systems, it is also important to concomitantly redevelop the foundational bases of the economics discipline from better first principles. A key open question in the development of realistic micro-foundations is being able to account for individual differences in human preferences in ways that are predictive and grounded in scientific principles, a need that is met, in part, by the contributions made in this dissertation.

Thus, a more speculative future direction for this work involves aggregating predictions from micro-agents in simulated economic systems to develop macroeconomc models that predict stylized facts about the economy better than existing methods.

Such an effort could be interpreted as attempting to replicate the agent-based modeling program with agents that use belief update rules in line with scientific expectations, instead of ad hoc postulates common in current modeling practice. The critical missing piece for such a project to arrive at fruition is the absence of strong empirical and theoretical constraints on the form of the micro-macro aggregation, i.e. the means by which micro-level predictions are combined to yield macro-level descriptons. While DSGE methods address such questions in detail, it remains to be seen if their predictions are in line with the cognitive expectations of social decision models such as ours.

7.2.4 Scientific study of intrinsic motivation

We have briefly touched upon the ways in which this work relates to research on computational modeling of intrinsic motivations. We have not addressed this question in greater detail because we believe, along Fristonian lines [11] that the distinction between extrinsic and intrinsic sources of motivation is fallacious, and has led modeling efforts in this area considerably astray. Further establishing connections between our theory and existing research on intrinsic motivation is necessary to allow the transfer of insights obtained in our work to robotics and other AI applications.

Such applications, however, are not the main reason why we believe this avenue of research to be of great significance. The bigger reason arises from our sense of the evolutionary trajectory of the world's socio-economic system. Whether it arises out of resource constraints, or out of increasingly sophisticated automation, it appears likely that Bertrand Russell's prediction of the end of work[1] is likely to come true at some point in the future (quite possibly within our lifetimes). As many are discovering to their discomfort, reduction in the labor force necessary to sustain the economic infrastructure spells endemic unemployment and social unrest, irrespective of the political system. If productivity gains and/or limits to growth will, singly or together, continue to perpetuate massive structural unemployment for the foreseeable future, it becomes all the more important for both policy-makers and independent citizens to make an effort to understand how humans can entertain themselves without having to be externally rewarded. When Pascal said,

[1] Best expressed in his splendid essay, 'In praise of idleness'

> All of man's misfortune comes from one thing, which is not knowing how to sit quietly in a room,

in his *Pensees*, he did not quite foresee its applicability in this present sense. Nonetheless, as labor requirements drop over the coming decades, the ability to sit quietly in a room will become increasingly important as an existential requirement. Understanding the cognitive mechanisms that promote satisfaction through intrinsic factors is likely to figure prominently in societal efforts to inculcate this ability. We believe, therefore, that the scientific study of intrinsic motivation is a pressing societal need, and feel that extending our theory in this direction will yield concrete contributions to this effort.

7.2.5 Clarifying the role of dopamine in human decisions

A recurring theme throughout this dissertation has been the role of neurobiological evidence in constraining the form that possible theories of behavior can take, and affirmative demonstrations of the commensurability of our own proposal with such observations. For instance, in Chapter 2, we showed that the existence of comparison coding in the orbito-frontal neurons of monkeys precluded solely absolute representations of value in primate brain circuits [29]. Likewise, in Chapter 3, we appealed, in part, to empirical results collected by [82] to motivate the specific form of memory accessibilty, as well as to data from [79] relating our postulated mechanism of risk sensitivity with their results.

In future work, I hope to discover ways in which our predictions can be tested at the level of neurobiological experimentation. Several studies, indirectly supportive of our conclusions, have been emerging from the neuroeconomics community in the recent past. To take a specific example, considerable literature has emerged in trying to elucidate the role of dopamine in human decision-making. As Friston et al. [154] masterfully summarize, "Dopamine has been implicated in a bewildering variety of processes and pathologies in the human brain; ranging from cortical excitability to attentional deficits; from motor control to akinesia and set switching deficits in Parkinsons disease; from working memory to schizophrenia ; from reinforcement learning to addiction; from executive function to age-related cognitive decline; from reward prediction to failures of incentive salience; from exploration to psychomotor poverty."

While this remains speculation at this point, we believe that associating dopamine levels with predictive confidence resolves the problem of its multifarious uses, and generates a simple explanation for its function at different levels in the neurocognitive hierarchy. Our confidence is somewhat buttressed in this assertion by Friston's recent theoretical note that, proposing a role for dopamine identical to that assigned to predictive confidence for our theory's belief-update, clarifies a number of the functional connections mentioned above with an elegant model [154]. We note additionally that other researchers have conclusively shown that dopaminergic activity correlates strongly with prediction errors made by learning agents [26, 105, 155, 156] as well as with policy uncertainty [110], a role that is clearly compatible with our definition of confidence as a measure of cumulative surprise and uncertainty. Pulling these threads of evidence together, we feel that extending Friston's ideas by connecting dopaminergic activity with our definition of predictive confidence appears to be a straightforward and very exciting avenue for future work.

7.2.6 Deepening phenomenological implications

In this last section of the last chapter of this dissertation, I address what is, to me, the most important aspect of the research that I have conducted through the course of this dissertation, and hope to pursue in the future. It is my belief that progress in understanding behavior has been frequently retarded by an unwarranted pessimism about our ability to find deep insights into human nature - a concern that I feel has a partially theological origin. By treating the human mind smply as a physical entity, shaped by evolutionary adaptation to promote efficiency in processing beliefs, we have made the smplest and most natural assumption possible about a biological organ, with the results described in the foregoing pages.

Ultimately, however, this physicalist description of mind will not be complete until it can account for all possible phenomenologically accessible aspects of mind. The current proposal is asserted to explain the formation of preferences for different affordances in the world. It cannot yet explain vigor of response, nor the dynamics of switching between multiple tasks. Neither can it yet explain the deepest phenomenological aspects of human nature, viz. self-awareness, sense of agency and sense of anxiety. I believe that further deepening the model I have developed in this thesis to account for the possibility

of multiple tasks, and generalizing my conceptualization of predictive confidence to be computed over memory traces from multiple task-affordances will give us answers about these deep parameters of the mind. Specifically, I speculate that, just as cognitive effort minimization can be construed as a principle of least action explaining human behavior, we will ultimately be able to develop a conservation principle, implicating conservation of sense of agency as the deeper principle that more fully explains human behavior across multiple tasks and decisions. Of all the possible future directions of the work developed in this thesis, this is the one that I look forward to with the greatest anticipation.

7.3 Epilogue

The Road goes ever on and on
Down from the door where it began.
Now far ahead the Road has gone,
And I must follow, if I can,
Though quiet woods and silent hill
Have soothed an often troubled mind,
The feet, they yearn to wander still,
The Road for long will yet unwind.

References

[1] G. Gigerenzer and D. Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

[2] J.S. Mill. Essays on some unsettled questions of political economy, par 5.38, 1874.

[3] Neil Stewart, Nick Chater, and Gordon D.A. Brown. Decision by sampling. *Cognitive Psychology*, 53(1):1 – 26, 2006.

[4] A. Tversky and D. Kahneman. Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232, 1973.

[5] K. Friston. The free-energy principle: a unified brain theory? *Nat Rev Neuroscience*, 11:127–38, 2010.

[6] S. Gershman, , and R. Wilson. The neural costs of optimal control. In *Proceedings of Advances in Neural information processing systems (NIPS) 22*, 2010.

[7] Nisheeth Srivastava and Paul Schrater. An evolutionarily motivated model of decision-making under uncertainty. Available at SSRN: http://ssrn.com/abstract=1687205, 2010.

[8] A.G. Barto and R.S. Sutton. *Reinforcement Learning: an introduction*. Univesity of Cambridge Press, 1998.

[9] C. Camerer and T-H. Ho. Experienced-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

[10] Hubert L. Dreyfus. *What computers still can't do: a critique of artificial reason*. MIT Press, Cambridge, MA, USA, 1992.

[11] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293 – 301, 2009.

[12] Karl Friston, Chris Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3(130), 2012.

[13] Martin Heidegger and Joan Stambaugh. *Being and Time: A translation of Sein und Zeit*. State University of New York Press, 1996.

[14] Nisheeth Srivastava and Paul R. Schrater. Rational inference of relative preferences. In *Advances in Neural Information Processing Systems, NIPS*, 2012.

[15] Ido Erev, Eyal Ert, and Alvin E. Roth. A choice prediction competition for market entry games: An introduction. *Games*, 1(2):117–136, 2010.

[16] Nisheeth Srivastava and Paul Schrater. A value-relativistic decision theory predicts known biases in human preferences. In *Proceedings of the* 33^{rd} *Annual Meeting of the Cognitive Sciences Society (to appear)*, 2011.

[17] Nisheeth Srivastava, Komal Kapoor, and Paul Schrater. A cognitive basis for theories of intrinsic motivation. In *Proceedings of the First joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2011.

[18] Nisheeth Srivastava and Paul Schrater. Cognitive efficiency as a causal mechanism for social preferences. In *Proceedings of SocialCom/PASSAT*, pages 647–651, 2011.

[19] D. Kreps. *A Course in Microeconomic Theory*, pages 17–69. Princeton University Press, 1990.

[20] Nathaniel Daw and Michael J. Frank. Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, 113(3):259 – 261, 2009. Reinforcement learning and higher cognition.

[21] D. Kahneman. Perception, action and utility: the tangled skein. In M. Rabinovich, K. Friston, and P. Varona, editors, *Principles of Brain Dynamics: Global State Interactions*. MIT Pres, 2012.

[22] M. Rabin. Psychology and economics. *Journal of Economic Literature*, 36(1):pp. 11–46, 1998.

[23] R. D. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. Wiley, New York, 1957.

[24] J.R. Busemeyer and J.T. Townsend. Decision field theory: A dynamic cognition approach to decision making. *Psychological Review*, 100:432–459, 1993.

[25] A. Tversky and I. Simonson. Context-dependent preferences. *Management Science*, 39(10):pp. 1179–1189, 1993.

[26] W. Schultz, P. Dayan, and P. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

[27] P. Read Montague, Brooks King-Casas, and Jonathan D. Cohen. Imaging valuation models in human choice. *Annual Review of Neuroscience*, 29(1):417–448, 2006.

[28] I. Vlaev, N. Chater, N. Stewart, and G. Brown. Does the brain calculate value? *Trends in Cognitive Sciences*, 15(11):546 – 554, 2011.

[29] L. Tremblay and W. Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398:704–708, 1999.

[30] R. Elliott, Z. Agnew, and J. F. W. Deakin. Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, 27(9):2213–2218, 2008.

[31] Takayuki Hosokawa, Keichiro Kato, Masato Inoue, and Akichika Mikami. Neurons in the macaque orbitofrontal cortex code relative preference of both rewarding and aversive outcomes. *Neuroscience Research*, 57(3):434 – 445, 2007.

[32] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

[33] J.A. Gray and N. McNaughton. *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. Oxford University Press, 2000.

[34] J.v. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[35] L.J. Savage. *The foundations of statistics*. Wiley Press, New York, 1954.

[36] Robert Bordley and Marco LiCalzi. Decision analysis using targets instead of utility functions. *Decisions in Economics and Finance*, 23(1):53–74, 2000.

[37] J. R. Busemeyer, R. Barkan, S. Mehta, and A. Chaturvedi. Context effects and models of preferential choice: implications for consumer behavior. *Marketing Theory*, 7(1):39–58, 2007.

[38] Louisa C. Egan, Laurie R. Santos, and Paul Bloom. The origins of cognitive dissonance. *Psychological Science*, 18(11):978–983, 2007.

[39] I. Vlaev, B. Seymour, R.J. Dolan, and N. Chater. The price of pain and the value of suffering. *Psychological Science*, 20(3):309–317, 2009.

[40] L. Gabora and D. Aerts. Contextualizing concepts using a mathematical generalization of the quantum formalism. *Joural of Experimental and Theoretical Artificial Intelligence*, 14(4):327–358, 2002.

[41] W. Leong and D. Hensher. Embedding decision heuristics in discrete choice models: A review. *Transport Reviews*, 32(3):313–331, 2012.

[42] Nick Chater. Rational and mechanistic perspectives on reinforcement learning. *Cognition*, 113(3):350 – 364, 2009. Reinforcement learning and higher cognition.

[43] Neil Stewart. Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology*, 62(6):1041–1062, 2009.

[44] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98:1325–1348, 1990.

[45] Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutstrm. Estimating risk attitudes in denmark: A field experiment. *Scandinavian Journal of Economics*, 109(2):341–368, 2007.

[46] Steffen Andersen, GlennW. Harrison, MortenIgel Lau, and E.Elisabet Rutstrm. Elicitation using multiple price list formats. *Experimental Economics*, 9:383–405, 2006.

[47] H. Zhang and L. Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6, 2012.

[48] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

[49] U. Chajewska, D. Koller, and D. Ormoneit. Learning an agent's utility function by observing behavior. In *ICML*, pages 35–42, 2001.

[50] C.G. Lucas, T. Griffiths, F. Xu, and C. Fawcett. A rational model of preference learning and choice prediction by children. In *NIPS*, pages 985–992, 2008.

[51] A. Jern, C. Lucas, and C. Kemp. Evaluating the inverse decision-making approach to preference learning. In *NIPS*, pages 2276–2284, 2011.

[52] D. Hensher, J. Rose, and W. Greene. *Applied Choice Analysis: A Primer*. Cambridge University Press, 2005.

[53] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. MIT Press, 1998.

[54] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, 2000.

[55] L. Shi and T. Griffiths. Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1669–1677. 2009.

[56] K. Canini, M. Shashkov, and T. Griffiths. Modeling transfer learning in human categorization with the hierarchical dirichlet process. In *ICML*, pages 151–158, 2010.

[57] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

[58] Stephen V. Burks, Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini. Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, 106(19):7745–7750, 2009.

[59] Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–60, 2010.

[60] Boyle P. A., Yu L., Buchman A. S., Laibson D. I., and Bennett D. A. Cognitive function is associated with risk aversion in community-based older persons. *BMC Geriatrics*, 11, 2011.

[61] Thomas L. Griffiths and Joshua B. Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773, 2006.

[62] Ido Erev, Ira Glozman, and Ralph Hertwig. What impacts the impact of rare events. *Journal of Risk and Uncertainty*, 36:153–177, 2008.

[63] Richard Gonzalez and George Wu. On the shape of the probability weighting function. *Cognitive Psychology*, 38(1):129 – 166, 1999.

[64] John F. Anderson. Act: A simple theory of complex cognition. *American Psychologist*, 51(4):355–365, 1996.

[65] John E. Laird. Extending the soar cognitive architecture. In *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pages 224–235. IOS Press, 2008.

[66] Ron Sun. Cognitive social simulation incorporating cognitive architectures. *Intelligent Systems, IEEE*, 22(5):33 –39, sept.-oct. 2007.

[67] R. Laird, J. Newell, and P. Allen. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[68] Daniel; Byrne Michael D.; Douglass Scott; Lebiere Christian; Qin Yulin Anderson, John R.; Bothell. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

[69] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[70] Andrew R.A. Conway, Michael J. Kane, and Randall W. Engle. Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, 7(12):547 – 552, 2003.

[71] Keisuke Fukuda, Edward Vogel, Ulrich Mayr, and Edward Awh. Quantity not quality: The relationship between fluid intelligence and working memory capacity. *Psychon Bull Rev.*, 17, 2010.

[72] C. Gonzalez and V. Dutt. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118:523–551, 2011.

[73] M. Rabin and G. Weizsacker. Narrow bracketing and dominated choices. *American Economic Review*, 99(4):1508–43, 2009.

[74] S. McClure, D. Laibson, G. Loewenstein, and J. Cohen. Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695):503–507, 2004.

[75] Ming Hsu, Ian Krajbich, Chen Zhao, and Colin F. Camerer. Neural response to reward anticipation under risk is nonlinear in probabilities. *The Journal of Neuroscience*, 29(7):2231–2237, 2009.

[76] B. Shiv and A. Fedorikhin. Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, 26(3):278–92, 1999.

[77] Daniel J. Benjamin, Sebastian Andres Brown, and Jesse M. Shapiro. Who is behavioral? cognitive ability and anomalous preferences. Levine's working paper archive, David K. Levine, 2006.

[78] Drazen Prelec. The probability weighting function. *Econometrica*, 66(3):497–528, May 1998.

[79] Martin P. Paulus and Lawrence R. Frank. Anterior cingulate activity modulates nonlinear decision weight function of uncertain prospects. *NeuroImage*, 2005.

[80] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[81] S. F. Taylor, N. Tishby, and W. Bialek. Information and fitness. *ArXiv e-prints*, December 2007, 0712.4382.

[82] I. Nevo and I. Erev. On surprise, change and the effect of recent outcomes. *Frontiers in Psychology*, 3(24), 2012.

[83] Matilde Bombardini and Francesco Trebbi. Risk aversion and expected utility theory: A field experiment with large and small stakes, 2005.

[84] Dan Ariely. *Predictably irrational: The Hidden Forces That Shape Our Decisions*. Harper Collins, 2009.

[85] S. Ayal and G. Hochman. Ignorance or integration: The cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making*, 22:455–474, 2009.

[86] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329 – 349, 2009. Reinforcement learning and higher cognition.

[87] J. Persky. Retrospectives: The ethology of homo economicus. *The Journal of Economic Perspectives*, 9(2):221–231, 1995.

[88] P.A. van der Helm and E.L.J. Leeuwenberg. Accessibility, a criterion for regularity and hierarchy in visual pattern codes. *Journal of Mathematical Psychology*, 35:151–213, 1991.

[89] P. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[90] H. Simon. *Models of man: social and rational*. Wiley Press, 1957.

[91] J. Conlisk. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.

[92] S.J. Russell. Rationality and intelligence. *Artificial Intelligence*, 94:57–77, 1997.

[93] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[94] Jurgen Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, 2002.

[95] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley Press, 1978.

[96] Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.

[97] J. Klayman and Y.-W. Ha. Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94:211–228, 1987.

[98] P.C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12:129–140, 1960.

[99] S. Deese and J. Kaufman. Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54:180–187, 1957.

[100] M. Glanzer and A.R. Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behaviour*, 5:351–60, 1966.

[101] P.B. Sederberg, M.W. Howard, and M.J. Kahana. A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115:893–912, 2008.

[102] D. Kahneman. Objective happiness. In D. Kahneman, E. Diener, and N. Schwarz, editors, *Well-Being: The Foundations of Hedonic Psychology*, pages 3–25. Russell Sage, 1999.

[103] J. Pascual-Leone. A mathematical model for the transition rule in piaget's developmental stages. *Acta Psychologica*, 32:1704–1711, 1970.

[104] B.F. Skinner. *Beyond freedom and dignity*. Knopf, 1971.

[105] PR Montague, Peter Dayan, and TJ Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16:1936–47, 1996.

[106] Hubert L. Dreyfus. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, 171(18):1137–1160, 2007.

[107] R. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[108] J. Metcalfe and A. P. Shimamura. *Metacognition: knowing about knowing*. MIT Press, 1994.

[109] Nathaniel Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711, 2005.

[110] Quentin J.M. Huys and Peter Dayan. A bayesian formulation of behavioral control. *Cognition*, 113(3):314 – 328, 2009. Reinforcement learning and higher cognition.

[111] S. Singh, A. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Proceedings of Advances in Neural information processing systems (NIPS) 17*, 2005.

[112] J. Anderson, D. Bothell, C. Lebiere, and M. Matessa. An integrated theory of list memory. *Journal of Memory and Language*, 38:341–380, 1998.

[113] J. Dickhaut, A. Rustichini, and V. Smith. A neuroeconomic theory of the decision process. *PNAS*, 106(52):22145–22150, 2009.

[114] D. Berlyne. *Conflict, Arousal and Curiosity*. McGraw-Hill, 1960.

[115] J.C. Horvitz. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96(4):651–656, 2000.

[116] M. Czikszenthmihalyi. *Flow - the Psychology of Optimal Experience*. Harper Perennial, 1991.

[117] P.-Y. Oudeyer, F. Kaplan, and V.V. Hafner. Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation, IEEE Transactions on*, 11(2):265 –286, 2007.

[118] O. Simsek and A. Barto. An intrinsic reward mechanism for efficient exploration. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML)*, 2006.

[119] J. Tooby and L. Cosmides. The psychological foundations of culture. In J. Barkow, L. Cosmides, and J. Tooby, editors, *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press, 1992.

[120] J. Anderson. *The Adaptive character of thought*. Erlbaum Press, 1990.

[121] Jonathan Sorg, Satinder Singh, and Richard L. Lewis. Internal rewards mitigate agent boundedness. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[122] Satinder Singh, Richard L. Lewis, Andrew G. Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. In *IEEE Transactions on Autonomous Mental Development*, 2010.

[123] A. Rubinstein. *Modeling bounded rationality*. Prentice-Hall, 2003.

[124] Satinder Singh, Richard L. Lewis, and Andrew G. Barto. Where do rewards come from? In *Annual Conference of the Cognitive Science Society (CogSci)*, 2009.

[125] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *International Conference on Machine Learning*, pages 697–704, 2006.

[126] Jonathan Rubin, Israel Nelken, and Naftali Tishby. Cortical representation of subjective surprise. In *Preceedings of Cosyne*. Nature, 2011.

[127] G.A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. cued/f-infeng/tr 166, cambridge university, 1994.

[128] Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

[129] G. F. Loewenstein, L. Thompson, and M. H. Baserman. Social utility and decision making in interpersonal contexts. *J. Pers. Soc. Psychol.*, 57:426–441, 1989.

[130] E. Fehr and U. Fischbacher. The nature of human altruism. *Nature*, 425(6960):785–791, 2003.

[131] A. Falk, E. Fehr, and U. Fischbacher. Testing theories of fairness[mdash]intentions matter. *Games Econ. Behav.*, 62:287–303, 2008.

[132] J. Andreoni and B. Bernheim. Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. *Econometrica*, 77:1607–1636, 2009.

[133] Stefan T. Trautmann. A procedural extension of the fehr-schmidt model of inequality aversion, M Phil thesis, university of amsterdam, 2006.

[134] E. Fehr and K.M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114:817–868, 1999.

[135] Douglas R. Hofstadter. *Metamagical Themas: questing for the essence of mind and pattern.* Bantan Dell, 1985.

[136] R. Axelrod. *The Evolution of Cooperation.* Basic Books, 1984.

[137] Robert Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, 1971.

[138] David Kraines and Vivian Kraines. Learning to cooperate with pavlov: an adaptive strategy for the iterated prisoner's dilemma with noise. *Theory and Decision*, 35:107–150, 1993.

[139] Martin Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364:56–58, 1993.

[140] James K. Rilling, David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, and Clinton D. Kilts. A neural basis for social cooperation. *Neuron*, 35(2):395 – 405, 2002.

[141] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[142] K Carley. A theory of group stability. *American Sociological Review*, 56(3):331–354, 1991.

[143] S. M. Kosslyn. On the evolution of human motivation: The role of social prosthetic systems. In S. M. Platek, T. K. Shackelford, and J. P. Keenan, editors, *Evolutionary cognitive neuroscience*. MIT Press, 2006.

[144] H B Kaplan, R J Johnson, and C A Bailey. Deviant peers and deviant-behavior - further elaboration of a model. *Social Psychology Quarterly*, 50(3):277–284, 1987.

[145] R. Albert and A-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[146] G.L. Robins, T.A.B. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29:192–215, 2007.

[147] A. Giddens. *Constitution of society: Outline of the theory of structuration*. University of California Press, 1986.

[148] J. Broekens, W. A. Kosters, and F. J. Verbeek. Affect, anticipation, and adaptation: Affect-controlled selection of anticipatory simulation in artificial adaptive agents. *Adaptive behavior*, 15:397–422, 2007.

[149] Ron Sun. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, New York, NY, USA, 1 edition, 2008.

[150] G.A. Akerlof. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84:488–500, 1970.

[151] Robert Lucas. Econometric policy evaluation: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, pages 19–46, 1976.

[152] N. Kaldor. The irrelevance of equilibrium economics. *Economic Journal*, 82:1237–1255, 1972.

[153] Paul Krugman. How did economists get it so wrong? new york times, 2009.

[154] Karl J. Friston, Tamara Shiner, Thomas FitzGerald, Joseph M. Galea, Rick Adams, Harriet Brown, Raymond J. Dolan, Rosalyn Moran, Klaas Enno Stephan, and Sven Bestmann. Dopamine, affordance and active inference. *PLoS Comput Biol*, 8(1):e1002327, 01 2012.

[155] Y. Niv, N. D. Daw, and P. Dayan. Choice values. *Nature Neuroscience*, 9(8):987–988, 2006.

[156] Y. Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.