UNIVERSITY OF CALIFORNIA,
IRVINE

PROJECTIBILITY AND THE SELF-ASSEMBLY OF LEARNING

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Christian Torsell

Dissertation Committee:
Distinguished Professor Jeffrey A. Barrett, Chair
Distinguished Professor Brian Skyrms
Chancellor's Professor Cailin O'Connor
Chancellor's Professor Simon Huttegger
Professor Kenny Easwaran

2026

DEDICATION

To my parents

and to the memory of my grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This project would not have been possible without the guidance and support of my dissertation committee: Jeff Barrett, Brian Skyrms, Cailin O'Connor, Simon Huttegger, and Kenny Easwaran. Each has provided valuable input on my work and offered an instructive example of philosophical depth and clarity in their own. Jeff deserves special thanks for his patience and generosity over the many hours we spent in his office discussing the ideas that grew into this dissertation. In these conversations, he modeled a compelling philosophical sensibility combining thoroughgoing pragmatism, broad curiosity, and a commitment to scientific fluency. I hope it's rubbed off.

Each chapter of the dissertation has been shaped by discussions in the Social Dynamics Seminar and the Game Theory Lab at UC Irvine. I am grateful to participants in both groups, and to Brian Skyrms, Louis Narens, Simon Huttegger, and Jeff Barrett for their roles in facilitating them. Thanks is due as well to an audience at the 2025 Philosophy, Politics, and Economics Society Annual Meeting in New Orleans who provided helpful feedback on an early version of the project that became the second chapter of the thesis. The third chapter of the dissertation benefited from discussion at the 2024 Workshop on Self-Assembling Games at UC Irvine. Hannah Rubin was among the among the workshop participants, and she deserves thanks both for several helpful conversations and for setting me on the path to Irvine in her spring 2019 game theory course at Notre Dame.

Portions of the research presented in Chapter 3 also appear in Christian Torsell and Jeffrey A. Barrett, "Learning how to learn by self-tuning reinforcement," Synthese 203, 209 (2024). Used with permission from Springer Nature B.V. Jeffrey A. Barrett directed and supervised research which forms the basis for the dissertation.

I am grateful to many friends for their support throughout graduate school. Among current and former LPS students, these include Curtis Mason, Elijah Spiegel, True Gibson, Josiah Lopez-Wild, Jack VanDrunen, Daniel Herrmann, Ryan Chen, Rebecca Korf, Neil Crawford, Margaret Farrell, Matthew Coates, and Nathan Gabriel; outside the department, Patrick Kollman, Nick Cohen, and Eyob Zewdie deserve special mention. Daniel was particularly instrumental in helping me get on my feet when I entered the program. Curtis has been a constant companion, a wellspring of kindness, and a trusted euchre partner. The road has been less lonely for Patrick's steady friendship and the fun and camaraderie it's supplied in generous quantities.

I am also grateful to my friends and mentors on the Orange County/Los Angeles/Long Beach jazz scenes who have provided an invaluable creative outlet, most especially Joe Nguyen, Mathew Marsico, Johnny Martinez, Matthew Nelson, Ben Osborne, Mike Zilber, and Daniel Rotem. While their contribution to this project was indirect, it was nevertheless critical.

I owe more than I can say to my parents, Dan and Leslee Torsell. They, along with my grandparents, provided the indispensable foundation for my current trajectory. My siblings, Elli, Patrick, and Tony have been beloved companions along the way.

VITA

Christian Torsell

EDUCATION

Doctor of Philosophy in Philosophy **2026**
University of California, Irvine *Irvine, California*

Bachelor of Arts in Philosophy **2021**
University of Notre Dame *Notre Dame, Indiana*

PUBLICATIONS

Egalitarianism and Evolution **2026**
Theory and Decision

Task Switching and Projectibility **2026**
Synthese

Learning to Forget **2025**
British Journal for the Philosophy of Science; with Jeffrey A. Barrett

Learning in Crawford-Sobel Games **2025**
British Journal for the Philosophy of Science; with Jeffrey A. Barrett, Cailin O'Connor, and Brian Skyrms

Janina Hosiasson and the Value of Evidence **2024**
Studies in History and Philosophy of Science

Learning How to Learn (by Self-Tuning Reinforcement) **2024**
Synthese; with Jeffrey A. Barrett

ABSTRACT OF THE DISSERTATION

Projectibility and the Self-Assembly of Learning

By

Christian Torsell

Doctor of Philosophy in Philosophy

University of California, Irvine, 2026

Distinguished Professor Jeffrey A. Barrett, Chair

Issues of projectibility are central to the philosophy of induction. The basic problem goes like this. Inductive learning involves projecting regularities in past experience onto unobserved cases. But any body of experience can be described as exhibiting any number of regularities, and depending on which of these we project, induction will yield different beliefs and predictions. A principled procedure for identifying regularities that provide *good* guidance for belief and prediction—i.e., *projectible* ones—has proven elusive. Nevertheless, the impressive track record of human and animal learning suggests that nature has found ways to deliver us assumptions of projectibility that track stable, practically relevant regularities in many contexts. This dissertation uses computational models to show how simple natural adaptive processes might accomplish this critical task. Specifically, the goal is to explain how such assumptions might emerge from trial-and-error learning in kinds of problems that arise frequently in nature. Chapter 1 supplies philosophical background and motivation. Chapter 2 considers projectibility in the context of *signaling*, building on a game theoretic model due to Lewis (1969). Chapter 3 turns to *discrimination learning*, modeling the emergence of assumptions of projectibility in a setting based on classic primate learning experiments conducted by Harry Harlow. Chapter 4 concerns the special challenges of projectibility posed by *task switching* problems. Chapter 5 concludes.

Chapter 1

Projectibility in Nature

1.1 Introduction

Even experts sometimes struggle to distinguish the Gray Bat from the Indiana Bat. There are subtle morphological tells, but they are often difficult to identify without close inspection in the hand or the lab. Many species in the genus *Myotis*, to which they both belong, share equally vexing similarities, earning them a share in the label "little brown job" (LBJ), applied to many species of small, nondescript bats.

But these morphological similarities hide important behavioral differences, particularly in roosting patterns. Both Gray Bats and Indiana Bats hibernate in caves during the winter. But while Gray Bats remain in caves or cave-like structures year-round, Indiana Bats take up residence in forests during the warmer months.

Consider two scientists, N. and L., who share a newfound interest in *Myotis* behavior. They independently consult a common body of evidence: a collection of detailed photographs of individual bats, each labeled with the location at which that bat was observed to roost.

Labels do not identify the species of the subjects, but in fact, the photos are mostly of Indiana Bats (all of which were found roosting in forests) with a few Gray Bats (all of which were founding roosting in caves) mixed in. All of them were taken during June, which happens to be the current month.

N., who is rather new to all this, sees a homogeneous collection of small, brownish bats. He notes the locations on the labels and does his best to discern patterns in the kinds of structures these LBJs inhabit. L. has been in the bat business for a long time, and though she hasn't worked extensively with *Myotis*, she recognizes the two species represented in the data, and with some effort, she can pretty reliably distinguish the Indianas from the Grays.

Once both have finished reviewing the data, they are shown an unlabeled photo of a Gray Bat and instructed to predict the kind of roost it inhabits. N., recalling that most of the LBJs he saw were forest-dwellers, reports that the bat likely lives in a forest (though, he admits, he can't be sure, since a few such bats have also been observed inhabiting caves). When L. makes her prediction, she notices the shorter ears and subtly silvery fur, and so restricts her attention to the photos she tentatively identified as depicting Gray Bats. She remembers that all of those were labeled as cave-dwellers, and so she reports that the pictured bat can mostly likely be found in a cave. L., of course, is correct, and N. is incorrect.

Nothing very surprising happens in that story. But it contains a lesson: What we learn depends on the patterns we are looking for, not just the empirical content of our observations. N. and L. saw the same photographs and labels, but N. was looking for patterns in roosting behavior among LBJs, while L. was looking for patterns in roosting behavior among Gray Bats and, separately, among Indiana bats. Put differently, N. supposes bat behavior is uniform across the LBJ category, while L. expects uniformity conditional on narrower species membership.

In philosophy, the problem of picking out which regularities to attend to for the purposes of

learning and action has been influentially posed as a problem of *projectibility*. The guiding question has been some version of the following: in reasoning inductively, how can we ensure that the hypotheses we consider are projectible, in that they are genuinely confirmed by their instances? The ultimate goal is characterize valid inductive inference. The cases cited to motivate this version of the problem are unlike our LBJ case. They involve scientists contemplating hypotheses that involve strange, "gerrymandered" predicates. Indeed, one reason for beginning with the bat story is to illustrate how the basic problem highlighted by those cases can arise in a less exotic context. We will have more to say about this below.

This dissertation is about "non-exotic" problems of projectibility and the methods by which learners in nature might solve them. A central theme is that the special philosophical problems raised in connection with projectibility are of a piece with a more general problem that confronts all organisms that can profit from experience. The challenge of determining which regularities in one's experience to attend to in learning can be raised without bringing in artificial-looking predicates, and even without making reference to scientific reasoning. It arises with just as much urgency for more primitive learners as it does for scientists making explicit inferences in the field or in the lab.

To set the stage, we will call on three philosophers. David Hume will orient us methodologically and thematically. Nelson Goodman will give a canonical articulation of (a special case of) our central problem. And Patrick Suppes will show us how to make good on our Humean commitments in approaching Goodman's problem. Having cleared some philosophical ground, we will then introduce our positive program for investigating natural problems of projectibility. Along the way, we briefly discuss the notion of *self-assembly* and its relationship to learning, which will be a recurrent theme.

1.2 David Hume on Justification

David Hume considered whether our inductive beliefs and practices rest on a rational foundation, and he concluded that they do not. Hume begins Section IV of the *Enquiry Concerning Human Understanding* by highlighting that many of our empirical beliefs outstrip the contents of our immediate experience. These are inferred from further beliefs about causal relationships between events of different kinds, which we arrive at by generalizing from past experience. Experience discloses that *this* A-type event was followed by a B-type event, and *that* A-type event was followed a B-type event, and so were all these others, and we come to believe on that basis that A-type events are *invariably* followed by B-type events in virtue of a causal connection between them.

But, Hume points out, that B-type events have always been observed to follow A-type events does not entail that any *unobserved* A-type event will be followed by a B-type event. When it comes to matters of fact, it is always a logical possibility that future experience will upset a so-far-exceptionless pattern. Thus, "If we be ... engaged by arguments to put trust in past experience, and make it the standard of our future judgement, these arguments must be probable only, or such as regard matter of fact and real existence" (*Enquiry*, par. 30). But Hume has already told us that all our reasoning concerning matters of fact is "founded on" the relation of cause and effect, and our beliefs about causal relationships arise from projecting regularities in our past experience onto future experience—a kind of inference that "proceed[s] upon the supposition that the future will be conformable to the past" (Hume, 2007, 26). So, any probable argument with that supposition as its conclusion must be circular.

Hume's response to these skeptical doubts is not to propose an alternative method for forming beliefs about the world, one that rests on a securer foundation in reason. He does not advise that we weaken our practical commitment to induction at all. In fact, as we will see, Hume

thinks it would be impossible to follow such advice, since we cannot help but assume certain observed regularities will be carried into the future, given our psychological nature.

Instead, in his "Skeptical Solution of These Doubts," Hume considers how our inductive beliefs and practices are *naturally learned*. Here, he introduces the psychological mechanism of *custom*:

> [W]herever the repetition of any particular act or operation produces a propensity to renew the same act or operation, without being impelled by any reasoning or process of the understanding, we always say, that this propensity is the effect of Custom. (Hume, 2007, 32)

Hume identifies the unreasoned, automatic operation of custom as the mechanism that converts accumulated experience into the beliefs about matters of fact that guide our expectations and actions:

> It is that principle alone which renders our experience useful to us, and makes us expect, for the future, a similar train of events with those which have appeared in the past. Without the influence of custom, we should be entirely ignorant of every matter of fact beyond what is immediately present to the memory and senses. (Hume, 2007, 32)

The psychology of human inductive learning is thus analogous to that of "animals, who, by the proper application of rewards and punishments, may be taught any course of action" (Hume (2007), 77). What's more, Hume took its effects to be irresistible, as "unavoidable as to feel the passion of love, when we receive benefits; or hatred, when we meet with injuries" (Hume, 2007, 34). We *cannot help* but form beliefs by the operation of custom, given our natural psychology.

In the transition from his introduction of the "Skeptical Doubts" about induction to his

proposal of the "Skeptical Solution" highlighting its psychological irresistibility, Hume makes two significant moves.[1] First, he turns our attention from the normative epistemic status of our inductive practices to their natural psychological origins. Second, he foregrounds the role of inductive belief in guiding action. This is perhaps most evident in a passage that appears near the end of Section V of the *Enquiry*:

> Custom is that principle, by which this correspondence [between the course of nature and the succession of our ideas] has been effected; so necessary to the subsistence of our species, and the regulation of our conduct, in every circumstance and occurrence of human life. Had not the presence of an object, instantly excited the idea of those objects, commonly conjoined with it, all our knowledge must have been limited to the narrow sphere of our memory and senses; and we should never have been able to adjust means to ends, or employ our natural powers, either to the producing of good, or avoiding of evil. Those, who delight in the discovery and contemplation of final causes, have here ample subject to employ their wonder and admiration. (40)

Here, Hume maintains that learning by custom is sufficient to explain the role played by our inductive beliefs and expectations in producing successful action and prediction; rational justification is not necessary.

In his positive story about induction, Hume thus turns our attention from rational justification to natural learning. The Skeptical Solution aims to describe the psychological mechanisms responsible for our adoption and exercise of familiar inductive practices, not show that induction is justified.

Much that Hume says is useful in orienting our treatment of projectibility. We too will turn away from questions of rational justification and toward natural learning. And we will follow

[1] See Barrett (2024), Barrett (2025), and Torsell and Barrett (2024) for further elaboration of this interpretation of Hume.

Hume in his emphasis on the continuity between inductive learning in humans and in other animals. However, we have the benefit of approaching these problems after over 250 years of developments in the science of learning that Hume did not live to see. So we will use a somewhat different set of tools than he did. We will return to these issues shortly. But first we will consider a classic philosophical treatment of projectibility.

1.3 Nelson Goodman on Projectibility

Nelson Goodman took Hume's point about justification. Goodman began his exposition of the "new riddle of induction" by rehearsing the old problem, and he endorsed Hume's diagnosis:

> Come to think of it, what precisely would constitute the justification we seek? If the problem is to explain how we know that certain predictions will turn out to be correct, the sufficient answer is that we don't know any such thing. If the problem is to find some way of distinguishing antecedently between true and false predictions, we are asking for prevision rather than for philosophical explanation. (Goodman (1983), pp. 67-8)

But he wasn't satisfied with rehashing the negative half of Hume's story. To clear the way for progress on a positive response to Hume's challenge, Goodman urged that the problem be recast in terms of the validity of inductive inferences. On this approach, the goal is not to demonstrate that induction must produce true beliefs or accurate predictions, but rather to lay bare the rules we in fact make use of (perhaps implicitly) in evaluating inductive reasoning. Goodman understood these rules as a product of a process of reflective equilibration, in which the rules are judged by the acceptability of the inferences they license and inferences are judged by their compatibility with the provisionally-accepted rules:

> An inductive inference . . . is justified by conformity to general rules, and a general rule by conformity to accepted inductive inferences. Predictions are justified if they conform to valid canons of induction; and the canons are valid if they accurately codify accepted inductive practice.. . . The problem of induction is not a problem of demonstration but a problem of defining the difference between valid and invalid predictions. (Goodman (1983), pp. 67-8)

In posing his "new riddle of induction," Goodman's target was the view that valid inductive inference could be characterized purely syntactically. In a standard kind of inductive inference, a hypothesis positing a universal regularity (e.g., "All As are Bs") is taken to be confirmed by its instances (e.g. "This A is a B"), and each additional observed instance is taken to increase the credibility of that hypothesis. Goodman pointed out that whether an inference of this kind accords with "valid canons of induction" as fixed by "established inductive practice" (Goodman, 1983, 64) depends on more than the syntax of the underlying argument; the semantics for the predicates figuring in the hypothesis matters, too.

To illustrate, Goodman introduced the predicate "grue." "Grue" applies to an object observed before a future time t if and only if it is green and to an object not observed before t if and only if it is blue. Let H_1 stand for the hypothesis that all emeralds are green and H_2 stand for the hypothesis that all emeralds are grue. Since any instance of H_1 observed prior to t is also an instance of H_2, a learner who observes many green emeralds before t will take her observations to confirm H_2 if in generalizing from those observations she considers H_2 rather than H_1.

The argument with H_1 as its conclusion is sanctioned as inductively valid, but the argument with H_2 as its conclusion seems perverse. We call H_1 *projectible* to express that, according to standard inductive practice, it is genuinely confirmed by its instances. Since H_2 lacks this property, we call it *non-projectible*.

Our problems don't stop at "grue." We can cook up hypotheses invoking similarly gerrymandered predicates of an unlimited variety, with the consequence that our accumulated observations apparently confirm all sorts of strange hypotheses and the bizarre predictions they imply. Goodman puts the point memorably: "To say that valid predictions are those based on past regularities, without being able to say *which* regularities, is quite pointless. Regularities are where you find them and you can find them anywhere" (Goodman (1983), 81).

Goodman sought to identify a principled basis for distinguishing the projectible hypotheses from the non-projectible ones. His strategy considers the predicates that figure in candidate laws one-by-one. In particular, it appeals to the relative *entrenchment* of various predicates. To see what he had in mind, it will be helpful to introduce Goodman's notion of *actual projection*:

> A hypothesis will be said to be *actually projected* when it is adopted after some of its instances have been examined and determined to be true....Actual projection involves the overt, explicit formulation and adoption of the hypothesis—the actual prediction of the outcome of the examination of further cases. (87-88)

A hypothesis is actually projected when it is used in a particular, concrete case of prediction. A predicate's degree of *entrenchment* is an increasing function of the number of times it has figured in an actually (and successfully) projected hypothesis. If one predicate has appeared in more successful past projections than another, then the first is better entrenched than the second.

Goodman's solution takes the form of a rule for choosing among predicates that figure in conflicting projections, where two projections conflict if they disagree on at least one unexamined case. The rule is that the more entrenched predicate is to be preferred in cases of conflict. The desired result follows immediately: "green" is projectible, and "grue" is not,

because "green" has in fact been used in many more successful past projections than "grue." And so, in considering the collection of observations of emeralds, we should take them to confirm the green hypothesis, rather than the conflicting grue hypothesis.

Goodman's treatment of the "new riddle" vividly illustrates how the course of inductive learning depends on the categories we use in individuating observations. In this, it is instructive. However, as Patrick Suppes will show us, it fails to display the problem in its full generality. And there are modeling frameworks both more precise and more general than Goodman's entrenchment-based approach for explaining how learners in nature might learn inductive assumptions of projectibility. We turn to these issues in the next two sections.

1.4 Patrick Suppes on Learning

In "Learning and Projectibility," Patrick Suppes shows us how to think about the grue challenge in terms of Hume's central concerns. Suppes suggests a reframing of Goodman's problem motivated by two central ideas:

> First, much of learning is nonverbal, and consequently a statement-type formulation of inductive problems is inappropriate. Second, scientists are no different from other mammals trying to survive in the wilderness in terms of the basic problems of projectibility, or put in another standard way, prediction. (Suppes (1994), 263)

The complaint is that by framing the basic problem of projectibility as a problem of characterizing a kind of valid inference, the full scope of the problem Goodman raises is obscured. In particular, Goodman's treatment says nothing about problems of projectibility as they arise for non-linguistic learners. *All* learning involves selective responsiveness to patterns in experience, and on its own, experience itself is silent on which patterns are reliable guides

for action and prediction.

In highlighting the generality of the problem, Suppes turns our attention to learning as it occurs in nature:

> Artificial predicates like grue . . . are no problems for animals, because they are not considered, but projectibility is. Much learning occurs in relation to features of the environment that are partly stationary, and those stationary features have nice properties of projection. What is important about learning, however, is dynamic adaptability. . . . For animals that want to survive, there is a problem of projectibility, but it is not exactly Goodman's problem. Can what they have learned in the past be flexible enough to permit them to adapt to a changing environment in the future? (Suppes (1994), 264)

Animals inhabit environments that present a variety of recurrent choice situations and which include many variable features learners might attend to in distinguishing between these situations. Successful action is possible only if they attend to features that are informative about the conditions for practical success in the problems they frequently encounter, and if they do not overgeneralize what they learn in one setting such that they cannot adjust to changing conditions in the future.

If we accept Suppes' framing of the problem, what room is there for *normative* inquiry concerning projectibility? We've already eschewed traditional questions about the rational foundations of inductive methods. And if we abandon Goodman's concerns with validity, we can no longer invoke a set of established inference rules as an evaluative standard. Suppes' response is pragmatistic: different methods of learning (which embody different inductive assumptions concerning projectibility) may be evaluated by measuring their degree of predictive success in particular contexts. In explicating this suggestion, Suppes connects his approach to Hume's:

> A measure just of the blind success of predictions is limited, but here the measure is tied to individual mechanisms of learning. Beyond such mechanisms there is no meaningful concept of validity of predictions or deeper concepts of justification of induction. As Hume rightly said most learning in ordinary experience is unconscious and unreflective. There can be no logical or transcendental justification of the inherited biological mechanisms of learning beyond some measure of their success. (Suppes (1994), 271)

While we seek neither an ultimate rational foundation for our inductive practices nor a description of the rules we appeal to in sorting good inductive inferences from bad ones, we can still ask whether our inductive methods and practices are conducive to successful action and prediction in particular learning problems. For the Humean naturalist, this is enough.

We will follow Suppes in conceiving of projectibility as a problem for all learners. And we will adopt his emphasis on pragmatic success in assessing the forms of learning that appear the chapters that follow. Now, having been oriented by Hume, Goodman, and Suppes, we turn to a more detailed description of the our own project's aims and methods.

1.5 The Self-Assembly of Learning

In what follows, we use computational models to consider how simple learners might come to adopt inductive assumptions of projectibility that serve them well in problems that arise commonly in nature. In particular, we consider projectibility in the context of *signaling*, *discrimination learning*, and *task-switching*. Each corresponds to a setting in which successful learning requires selectively attending to patterns in experience that are informative about the conditions for successful action, where the relevant patterns aren't known in advance. And each poses a learning problem that is generic, in the sense that it, or some version of

it, confronts a wide array of natural learners of varying levels of cognitive and biological complexity.

Since our target is learning as it occurs in diverse kinds of natural agents and environments, we will want to invoke methods of learning that are biologically and psychologically plausible, even for relatively primitive organisms. Hume's notion of custom is on the right track. But we can avail ourselves of richer resources than Hume could in studying this sort of learning. On the empirical side, psychologists have produced detailed records of human and animal behavior in a dizzying variety of learning problems. And on the theoretical front, a broad array of formal learning rules have been developed and investigated, many of these sharing close affinities with Hume's notion of custom. We will help ourselves to the fruits of both kinds of enterprise in the models developed below.

The models of learning we draw on are forms of reinforcement learning in which an agent's choice dispositions are directly shaped by rewards and punishments. The leading idea these learning rules formalize was influentially articulated by American psychologist Edward Thorndike:

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. (Thorndike (1911), 244)

Thorndike called this the *law of effect*. The idea is that there is a random element in choice behavior, and the probability with which an agent chooses a given type of action depends on the history of associations between past actions of that kind and positive or

negative outcomes that closely followed those actions. Thorndike formulated this law after conducting a series of experiments on learning in cats, dogs, and chicks. This work was influential in shaping the behaviorism that would dominate American psychology in the first half of the twentieth century.

In psychology and, later, economics, a variety of mathematical models of learning formalizing Thorndike's basic idea have been developed.[2] These reinforcement learning models have a few significant virtues. There is strong empirical evidence that, in many contexts, both human and animal learning obeys the law of effect (Roth and Erev (1995); Erev and Roth (1998); Herrnstein (1970)). Another virtue is that these rules typically describe very simple patterns of adaptation that can be implemented with minimal cognitive machinery. What's more, even very primitive forms of reinforcement learning are efficacious in a large class of learning problems (see Beggs (2005), Argiento et al. (2009)).

The models we develop have the following structure. A learner begins with a collection of random dispositions over various kinds of responses (e.g., attentional and behavioral). These dispositions coevolve under a Thorndike-style reinforcement dynamics. As they do, novel structure gradually emerges in the composite learning system comprising the various dispositions, where the evolved structure reflects the learner's inductive assumptions concerning projectibility.

The endogenous emergence of new structure under a simple adaptive dynamics marks out the modeled phenomena as instances of *self-assembly*. Self-assembly occurs whenever an orderly system emerges, undesigned, from natural dynamical processes shaping the interactions of simpler systems that come to comprise it. Crystals, organisms (over evolutionary time), and many features of social organization self-assemble in this sense. Here, we are concerned with the self-assembly of learning systems.

[2]See, e.g., Bush and Mosteller (1951); Bush and Mosteller (1955); Roth and Erev (1995). The extremely influential learning model developed in Rescorla and Wagner (1972) is closely related to Bush and Mosteller's, though it was proposed as a model of classical, rather than operative, i.e. reward-based, conditioning.

Self-assembly has recently attracted significant attention in philosophy, particularly in the context of game theory (see Barrett (2025)). This literature elaborates the *theory of self-assembling games*, that is, games whose structure features coevolve with the strategies of their players. Although only one of the subsequent chapters deals with a setting immediately recognizable as a game, we will find the tools furnished by this framework useful in modeling the self-assembly of learning. Moreover, the single-agent models admit of an interpretation on which they describe a game of mutual interest played by a group of *subagents* who together comprise the individual learner, tightening the connection further.[3] We will have more to say about the theory of self-assembling games and its relationship to our project in the next chapter.

Having laid some philosophical groundwork and introduced the basic modeling framework in which we will work, we are ready to discuss the models.

[3]See O'Connor (2014) for an example of a game interpreted along these lines.

Chapter 2

Signaling

2.1 Projectibility in Signaling Games

We begin with projectibility in the context of signaling. Signaling is ubiquitous in nature, playing a critical role in animal communication, many basic cellular functions, and maintenance of homeostasis in larger-scale biological systems. In its most basic form, signaling involves two functional components: a *sender* which, by behaving differently in different states of its environment, transmits information about that environment to a *receiver*, which can observe and differentially respond to the sender's behavior but cannot observe the state directly. With respect to projectibility, the problem facing signalers in nature is that of identifying a basis for distinguishing states: When should a sender treat two discriminable states as the same for the purposes of learning and action, and when should it respond to them differently?

To get traction on this question, we will work with a popular game theoretic model of signaling, the *Lewis-Skyrms Game*. The Lewis-Skyrms game has its origin in David Lewis' 1969

doctoral dissertation, in which it is used to analyze the sense in which linguistic meaning can be said to be conventional (Lewis, 1969). In subsequent decades, it has come to be an important tool in studying biological communication and the evolution of language. Brian Skyrms' (2010, 2012) evolutionary analyses of this game and several variants have been particularly influential in shaping this literature, accounting for the second half of its hyphenated name. We will give a more complete description of the game in the next section.

While these game theoretic models have taught us a great deal about signaling, they are silent concerning the origins of the basis on which the players individuate environmental states. In particular, the Lewis-Skyrms setup assumes that the sender attends to just those differences in the environment that in fact make a difference for the players' payoffs from the start. Here, we introduce a novel variant of the Lewis-Skyrms game in which players must instead *learn* how to individuate states of nature. We consider how very simple kinds of reinforcement learning might enable players to learn successful strategies for distinguishing states, even as they learn a particular convention for communicating about those states.

2.2 Signaling, Self-Assembly, and Categorization

The theory of self-assembling games studies the emergence of structured strategic interactions from their more basic parts. Where standard evolutionary game theory models situations in which players' strategies in a fixed game evolve over time, the theory of self-assembling games allows features of the game itself, such as payoffs, strategy sets, and network structure, to evolve alongside players' strategies.

In extant work on self-assembling games, the Lewis-Skyrms signaling game and its variants have played a central role. The standard setup involves two players, a *sender* and a *receiver*, in an environment that can take on either of two states. The sender observes the state and

sends one of two signals to the receiver. The receiver observes the signal (but not the state) and chooses one of two available acts. Signals lack natural meaning; any association they come to have with states or acts must be fixed by convention. For each state, exactly one act yields a positive payoff for both the sender and the receiver. If the "wrong" act for the current state is chosen, then both receive a payoff of zero. More general models allow the number of states, signals, and acts to vary.

Evolutionary treatments of these games come in two varieties, depending on whether the dynamics is meant to describe evolution in a population or learning in individual players. The first models populations in which individuals are typed according to fixed sender or receiver strategies, given as mappings from states to signals or from signals to acts, respectively.[1] The proportion of the population comprising each type changes over time under a dynamics that specifies how type proportions grow or shrink depending on the payoffs players of that type receive in a given population state. The second kind of evolutionary treatment considers pairs of players whose strategies change over time under a learning dynamics.[2] In both kinds of model, the focus concerns which equilibria tend to be realized under the adaptive process under consideration.

Barrett and Skyrms (2017) extend these evolutionary treatments, showing how signaling games themselves might evolve from more basic games involving cue-reading and sensory manipulation. Subsequent work models other aspects of the self-assembly of signaling games. Herrmann and VanDrunen (2022), for example, consider how the conditional strategies available to the receiver might emerge endogenously, modeling a setting in which the receiver must *learn* which part of her environment constitutes the signal. Barrett (2023) shows how a signaling game might evolve out of the learned attentional and behavioral dispositions of a receiver and a pair of senders who must adopt distinct representational roles in order to achieve perfect signaling. Perhaps closest to our project is Barrett and LaCroix's (2022),

[1] See Hofbauer et al. (1998) for a textbook treatment.
[2] See Fudenberg and Levine (1998) for a textbook treatment.

which develops a model in which players in a signaling game who lack the expressive resources to distinguish all payoff-relevant states learn a signaling convention that partitions nature into equiprobable states. In these and related projects, the game itself emerges under the adaptive dynamics.

Here, we consider the self-assembly of the set of states that serve as the basis of the sender's conditional signaling dispositions. We investigate how the partition of states on which the sender conditions her choice of signal might emerge endogenously. Typically, signaling games are specified such that the states the sender distinguishes exactly track the conditions under which the receiver's action makes a difference to the players' payoffs. That is, the sender neither attends to payoff-irrelevant differences in his environment nor ignores payoff-relevant ones in choosing which signal to send. But in natural settings, organisms have no guarantee that the distinctions they attend to for the purposes of learning and action will correspond neatly to the conditions for practical success. Nature does not come out and tell them which perceptible differences correspond to pragmatic differences.

For natural learners, evolved perceptual saliences will often provide helpful guidance for individuating stimuli or contexts. When this is the case—most likely in frequently-encountered settings in which fitness-relevant goods are at stake—effective categorization is largely handled by innate psychological machinery. But sometimes learners find themselves in new practical contexts where evolved saliences offer misleading guidance or no guidance at all. In these settings, effective categorization requires *learning* which distinctions to attend to in individuating states. Here, we are concerned with this harder problem of categorization.

We consider how the sender in a signaling game might learn to individuate states of nature by means of a simple form of reinforcement learning. We first introduce a framework for modeling the emergence of the partition of states underlying the sender's conditional dispositions. We then use the framework to study a simple case that highlights the novel challenges that arise when no canonical categorization of states of nature is specified upfront. In our

model, before choosing which signal to send, the sender must choose which features of the environment she will condition her choice of signal on. Each possible choice corresponds to a partition of an underlying set of *atomic states*, where two atomic states are treated identically for the purposes of learning and action if they belong to the same element of the chosen partition. Some partitions can support the emergence of reliable communication, while others cannot. A central question guiding the analysis is: how often and under what conditions does the sender learn a state partition that can support perfect signaling?

As was indicated in the discussion of self-assembling games above, using signaling games to study the evolution of categorization is not new. The literature on *sim-max games*, a variant of the Lewis-skyrms game developed by Jäger (2007), provides another useful point of reference. In a sim-max game, a sender and a receiver aim to establish a convention for communicating about a state space structured by a metric defining similarity among states. Two states are similar to the extent that the receiver's possible actions have similar payoff consequences in each of them. O'Connor's (2014, 2015, 2019) work on these games illustrates their fruitfulness as tools for addressing canonical philosophical concerns about kinds and categorization.

Most directly relevant to our project is O'Connor (2014), which introduces a model of reinforcement learning in sim-max games. In O'Connor's model, learning proceeds along similar lines as in ours. Both the sender and the receiver are equipped with a set of conditional propensities: the sender's propensities concern which signal to send in each state, and the receiver's propensities concern which action to take conditional on each signal she might observe. These propensities, which are proportional to choice probabilities, are updated in response to success and failure in action. Whenever the receiver chooses an act that yields a positive payoff in the current state, the sender reinforces his conditional propensity to send the signal he just sent when in that state, and the receiver reinforces her propensity to choose the act she just chose conditional on observing that signal. Reinforcement magnitudes are

proportional to payoff. Categorization is relevant because there are more possible states than there are signals, and so the sender must learn which states to "lump together" by sending the same signal in each.

Our model differs from O'Connor's in two significant ways. First, as in the basic Lewis-Skyrms game, payoffs are binary rather than graded. While this choice reduces the model's generality, it comes with some practical advantages. While sim-max games are now entrenched in philosophy and game theory, they have not been as widely studied as the Lewis-Skryms game, particularly in connection with reinforcement learning. The tight formal connection between our model and standard Lewis-Skyrms games allows us to draw on the well-developed theory of the latter in illuminating the former. Moreover, working with a simple state space makes exposition more straightforward for a model that, as we will see, introduces novel complications elsewhere.

The second difference concerns the notions of categorization at work in O'Connor's and in our project. On her model (as in Barrett and LaCroix (2022)), distinct states count as belonging to the same evolved category if the sender learns to send the same signal in both states. However, the learning process by which the sender comes to respond identically to different states is one in which he always treats each state as distinct: for each state, he has a separate set of conditional signaling propensities that are activated and updated when he observes that state.

By contrast, in our model, the sender can learn to associate one and the same set of signaling propensities with *multiple* (atomic) states. Here, states are categorized together if they live in the same cell of the partition the sender uses to individuate states. If that partition lumps two atomic states together—call them state 1 and state 2—and the same signal A is optimal in both states, then in order to maximize reward, the sender need not separately learn the dispositions *if state is 1, send signal A* and *if state is 2, send signal A*; he instead need only learn one conditional disposition: *if state is 1 or 2, send signal A*. Since states 1 and 2 are

identified with the same cell of the partition, the sender activates and updates the same signaling propensities, whether he sees one or the other of those atomic states.

Introducing the possibility of associating multiple states with the same set of propensities makes a practical difference for learning. If two states respond identically to the receiver's actions in terms of payoff, then associating them with the same set of signaling propensities allows the sender to more quickly settle on a shared, determinate signal for that pair of states. Here, categorizing states together aids learning. It can also impair learning. If two states respond differently to the receiver's actions and the sender categorizes them together, the sender cannot learn to distinguish them in his signaling behavior, and so it is impossible to realize an optimal equilibrium in which the receiver is induced to choose the highest-payoff act in each state.

Treating categories in this way allows us to capture an important aspect of categorization that does not appear in extant treatments of category learning in signaling games: how we categorize stimuli determines how we generalize learning across contexts. In a review paper on concept learning in animals, Zentall et al. (2008) highlight the importance of this function of mental categorization for successful learning across many kinds of organisms:

> [B]eing able to sort objects, events, and relations into classes allows one to transfer learning to new stimuli or contexts that one judges to be perceptually, associatively, lexically, or functionally equivalent to those involved in the original learning. The advantage of this conceptual ability is that it provides great efficiency to learning. If a new environment can be identified as being similar to an old one, then prior learning can be applied, thereby reducing the costs and risks associated with new trial-and-error learning. (Zentall et al. (2008))

For categories to serve this purpose, it must be the case that two stimuli or contexts that are categorized together are *treated as (approximately) the same for the purpose of learning,*

such that what one learns in the course of engaging with the one is carried over to their engagements with the other. For reward-learning models like ours, this requires that the same set of action propensities be activated and updated in both contexts.

Returning to signaling, we are here interested not only in the categories reflected in the communicative convention learned by the sender and receiver, but also in the implicit categories guiding the sender's learning in the course of establishing that convention. In particular, we are interested in how those learning-guiding categories might *themselves* be learned. The next section describes the model in detail.

2.3 The Model

The *categorization game* generalizes the Lewis-Skyrms game by allowing the sender to choose how to categorize states of nature. This is done by introducing a distinction between the set of environmental differentia the sender could *in principle* attend to in individuating states, and those the sender *in fact* attends to.

The finest grain at which the sender can distinguish among states is given by the set of *atomic states* $\mathbf{A} = \{a_1, a_2, ..., a_n\}$. At each step t in the learning process, the state of nature observed by the sender is a random element of $\mathbf{A}$, written $S(t) = a$.

In identifying the state of the environment prior to sending a signal, the sender might lump together some atomic states, so that the distinctions she makes in identifying the state of nature correspond to a partition over $\mathbf{A}$ where some atoms belong to the same part. Call $\mathcal{S}_t$ the sender's *salience partition* at timestep t, where $\mathcal{S}_t = \{s_1^t, s_2^t, ..., s_m^t\}$ is a partition of $\mathbf{A}$. The sender's signaling dispositions at t, represented by a probability distribution over the available signals, are determined by which element of $\mathcal{S}_t$ the current atomic state belongs to. More precisely, given available signals $\sigma = \{\sigma_1, \sigma_2, ..., \sigma_k\}$, the probability with which a

given signal σ_i is sent at t can be written as $p_t(\sigma_i \mid \mathcal{S}_t(S(t)))$ (where $\mathcal{S}_t(S(t)))$ is the element of the salience partition in which the current atomic state lives). That is, we only need know which element of $\mathcal{S}_t$ the atomic state at t belongs to—not the atomic state itself—to know the sender's signaling dispositions.

On observing the signal, the receiver chooses an act from among $\alpha = \alpha_1, \alpha_2, ..., \alpha_k$. The payoff consequences of the receiver's act given the state $S(t)$ is dictated by a function $u : \mathbf{A} \times \alpha \rightarrow \mathbb{R}$. u takes a pair comprising an act and an atomic state and returns the pair of payoffs to the sender and receiver, respectively, given that state-act pair. Let u_i ($i \in \{sender, receiver\}$) map each act-state pair to the corresponding payoff for the player in role i. Use $\mathcal{R}$ to denote the partition sorting the atomic states into equivalence classes such that two states belong to the same class if they respond identically to the receiver's acts in terms of payoffs. Formally, $\mathcal{R}$ is the coarsest coarsening of $\mathcal{A}$ such that u is constant on its parts. Notice that, when $\mathcal{S}_t = \mathcal{R}$, the players' dispositions correspond to a standard Lewis-Skyrms signaling game where the sender's conditional signaling dispositions are assumed to reflect the underlying payoff structure from the get-go.

It may be helpful to consider a concrete example illustrating the roles of the various components of the model. Sarah and Rory sit on either side of a divider. In each of a series of trials, Sarah (but not Rory) is shown an image of either a circle or square which is either red or blue. She is equipped with two cards, one black and the other white. Upon observing the colored shape, she chooses one of these cards to slip under the divider to Rory's side, where Rory observes the color of the card and chooses which of two buttons, labeled "1" and "2," to press. Both players receive a small amount of money if Rory presses the "1" button when Sarah sees a blue shape or Rory presses the "2" button when Sarah sees a red shape; otherwise, neither gets anything. They are not informed of the conditions for success beforehand.

In this example, the atomic states are given by

$$\mathbf{A} = \{\text{red square, blue square, red circle, blue circle}\},$$

assuming Sarah can only distinguish among the stimulus arrays according to shape and color, and the reward partition is given by

$$\mathcal{R} = \{\{\text{red square, red circle}\}, \{\text{blue square, blue circle}\}\}.$$

The salience partition may vary from trial to trial depending on what feature(s) of the stimulus array Sarah notices and conditions her choice of card on. Suppose she initially assumes she is facing a shape discrimination task where color doesn't matter. In that case, her salience partition for the first trial would be written as

$$\mathcal{S}_1 = \{\{\text{red square, red circle}\}, \{\text{blue square, blue circle}\}\}.$$

Perhaps after many frustrating failures of coordination she comes to suspect the problem might be less straightforward, hypothesizing that what matters for the task at hand may be whether or not the object is a blue square. Then, we may write

$$\mathcal{S}_t = \{\{\text{red square, red circle, blue circle}\}, \{\text{blue square}\}\}.$$

Returning to the model, the learning process unfolds as follows. In a given timestep t, nature first selects the state a_t, where each possibility a_i is chosen with fixed probability. Then, the sender chooses a salience partition $\mathcal{S}_t$ by drawing from an urn containing balls labeled with each partition of $\mathbf{A}$. Call this the *partition urn*. She then observes which element of $\mathcal{S}_t$ contains $S(t)$. This observation determines which of his *category urns* the sender will draw from to determine which signal to send. She is equipped with one category urn for each

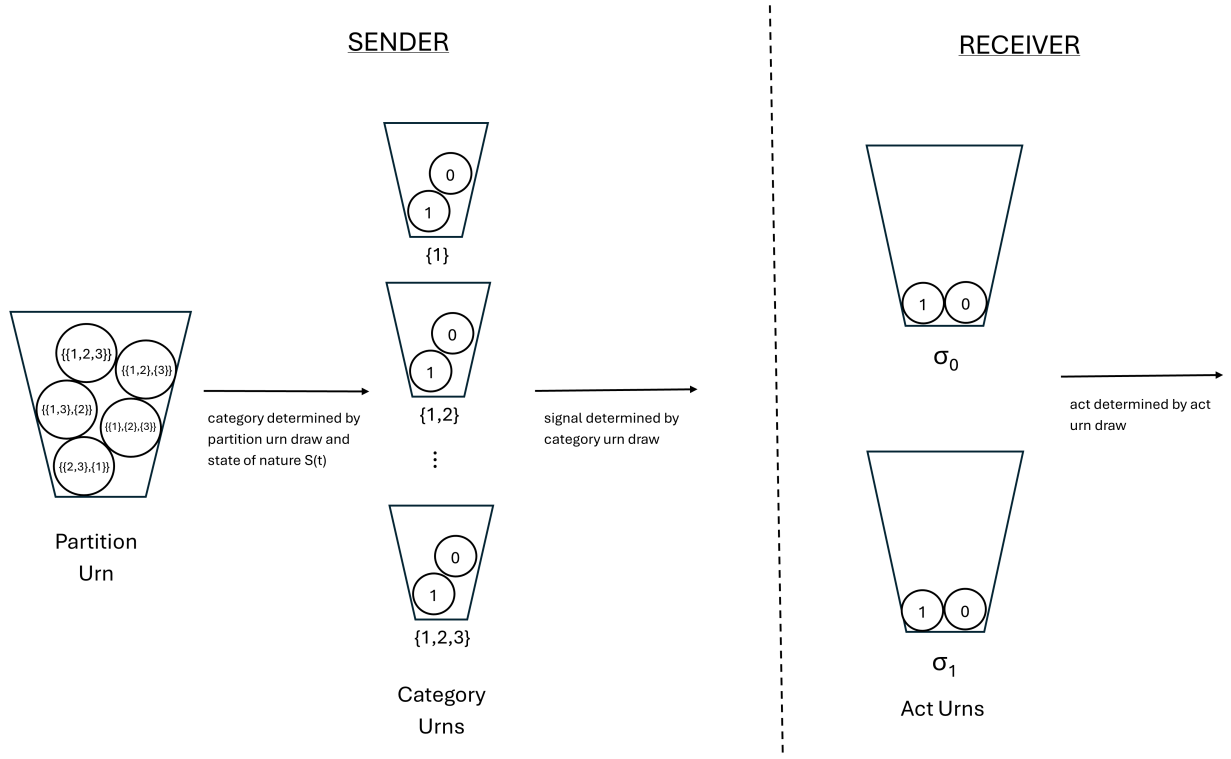

Figure 2.1: Initial urn contents for category game in which $S = \{1, 2, 3\}$, $\sigma = \{\sigma_0, \sigma_1\}$, $\alpha = \{\alpha_0, \alpha_1\}$.

subset of $\mathbf{A}$, so that she has determinate signaling dispositions for any partition-state pair. At t, she decides which signal to send by drawing from the category urn corresponding to $\mathcal{S}_t(S(t))$. Each category urn contains balls labeled $1, 2, ..., k$, corresponding to the signals $\sigma_1, \sigma_2, ..., \sigma_k$. She then sends the signal determined by her draw.

The receiver is equipped with k many urns, one for each signal he might receive. Each contains balls labeled $0, 1, ..., k-1$, corresponding to the acts $\alpha_0, \alpha_1, ..., \alpha_{k-1}$. On observing the signal, he draws from the corresponding urn to determine his act. If the play was successful, then both the sender and receiver note which kind of ball they drew from each urn and replace that ball, along with $r > 0$ many copies of the same kind of ball. This reinforces the disposition to make the corresponding response (choose the corresponding partition, send the corresponding signal ...) the next time the agent is in a relevantly similar situation. Call r the *reinforcement level*. If the play was unsuccessful, then both

players check the type of ball they drew from each urn and, as long as there are at least p many balls of that kind remaining in the urn, remove $p-1$ many balls of that type (without replacing the drawn ball). If there are fewer than p many balls of the relevant type remaining, then the player removes balls of that type until there is just one left. Call p the *punishment level*. Keeping at least one ball of each type in every urn ensures that no punishment episode drives a propensity to extinction, encouraging exploration throughout the learning process.

In the next section, we report results of computer simulations of learning in the categorization game. In analyzing those results, we will be guided by a few questions. Will the sender learn to use a state partition that can support perfect signaling? When she does, how often will it be the reward partition rather than some refinement of it which draws unnecessary distinctions? Will she typically learn a single canonical basis for individuating states, or will she mix over several partitions? How is learning to signal in the categorization similar to, or different from, learning to signal in the standard Lewis-Skyrms setup?

It is worth highlighting the interdependence between the sender's and the receiver's learning. The course of the sender's category learning depends on more than just her own choices. Whether a given partition gets reinforced depends on whether the sender's signaling dispositions conditional on using that partition are matched with appropriate *receiver* dispositions, so that when that partition is adopted, the signal sent is likely to induce the rewarded act. The partition of states is learned as the sender and receiver grope about for a signaling convention. This makes for a particularly challenging learning problem.

2.4 Simulations

Learning was simulated in categorization games with three, four, and five atomic states. To keep things manageable, we assume that, as above, there are just two acts and two states,

and that the reward partition divides the atomic states into two parts. Moreover, we assume state probabilities are uniform over the atomic states, and that players learn by $[+1, -1]$ reinforcement with punishment. For each set of parameters, we simulated 100 blocks of learning, each comprising 10^5 plays.

Before discussing simulation results, it will be helpful to introduce a bit more notation. Let $\mathbf{O}$ denote the set containing all refinements of the reward partition $\mathcal{R}$, including $\mathcal{R}$ itself, for a given categorization game.[3] Qualitatively, these are the partitions that make all the practically relevant distinctions among environmental states (and possibly more), and so can support perfect signaling. Since the games we study here involve many partitions that satisfy this requirement, we will track the players' degree of success in learning effective categorization schemes by recording the final probability with which the sender chooses any member of $\mathbf{O}$.

Since the labeling of signals, acts, and atomic states is arbitrary, we will call any two games G and G' *strategically equivalent* if, in addition to satisfying the assumptions above, their respective reward partitions $\mathcal{R}$ and $\mathcal{R}'$ divide the atomic states into an equal number of categories (i.e., $|\mathcal{R}| = |\mathcal{R}'|$), and for each category in $\mathcal{R}$ containing k many atomic states, there is a corresponding category in $\mathcal{R}'$ comprising k atomic states. For example, for categorization games with four atomic states 1, 2, 3, and 4, a version G in which $\mathcal{R} = \{\{1, 2\}, \{3, 4\}\}$ is strategically equivalent to another, G', with $\mathcal{R}' = \{\{1, 4\}, \{2, 3\}\}$. Both divide the possible environmental states into two categories containing two members each. The assumption that atomic state probabilities are uniform is crucial to our being able to treat these games as posing the same strategic problem to the players, since it follows from that assumption that any two sets of atomic states of equal cardinality occur with equal probability.

For a given categorization game G, we call the set comprising all the games that are strate-

[3] A partition $\mathcal{P}$ is a *refinement* of another partition $\mathcal{P}'$ if every element of $\mathcal{P}$ is a subset of an element of $\mathcal{P}'$.

gically equivalent to G a *strategic class*. Since we are restricting our attention to games in which there are two available acts and two corresponding elements of a game's reward partition, we can identify each strategic class with a pair of numbers, each giving the number of atoms in one part of the reward partition. Let k denote the total number of atomic states, with one part of the reward partition for a given game containing k_1 many atoms and the other part containing k_2 many atoms (so that $k_1 + k_2 = k$). We will use $\{k_1, k_2\}$ to denote the strategic class comprising all games whose reward partitions divide the space of atomic states into parts with k_1 and k_2 many atoms, respectively. If $k_1 = k_2$, then state probabilities are uniform over the reward partition. Otherwise, nature is biased in favor of one or the other of the parts.

prob. of most probable state	*corresponding strategic class*	*prop. of runs ending in pooling equilibrium*	*mean cumulative success rate*
0.8	$\{1, 4\}$	0.25	0.95
0.75	$\{1, 3\}$	0.14	0.964
0.67	$\{1, 2\}$	0.03	0.99
0.6	$\{2, 3\}$	0	1

Table 2.1: Simulation results for runs of 10^5 plays of $[+1, -1]$ reinforcement with punishment learning in $2 \times 2 \times 2$ Lewis-Skyrms games with biased state probabilities, averaged over 100 runs for each game.

If we identify states of nature with elements of the reward partition, each strategic class collapses to an ordinary Lewis-Skyrms game with probabilities over states of nature dictated by the number of atomic states in each element of the reward partition. Call this the *collapsed game* for that class of categorization games. The class of $\{1, 2\}$ categorization games corresponds to a $2 \times 2 \times 2$ Lewis-Skyrms game in which one state occurs with probability $\frac{1}{3}$ and the other with probability $\frac{2}{3}$. Moving up to four atomic states, the collapsed game for the $\{1, 3\}$ class is a Lewis-Skyrms game with state probabilities of 0.25 and 0.75, while the $\{2, 2\}$ class collapses to the vanilla case with unbiased nature. And so on. In addition to the usual signaling equilibrium, each collapsed game with non-uniform state probabilities has a *pooling equilibrium* in which the sender always sends just one signal and the receiver always

chooses the optimal act for the more probable state, for an expected success rate equal to that state's probability.

Thinking about collapsed games helps us see how the partition learning might function as a mechanism for the self-assembly of a signaling game. If the sender comes to reliably use the reward partition in a given categorization game, then the evolved game just is (an approximation of) its underlying collapsed game. If she instead learns to reliably use a different partition, then the evolved game can be thought of as a signaling game that deviates from the standard Lewis-Skyrms setup either by having multiple states that respond identically in terms of payoffs to the receiver's action (if the partition splits up states that share a part in the reward partition), or by having individual states that respond in a chancy fashion to the receiver's action (if the partition lumps together states that belonging to different parts of the reward partition). If the sender learns to mix among several different partitions, the evolved game has a more complicated structure that cannot be fully captured without explicitly including partition selection as part of the sender's strategy.

As a basis for comparison, table 1 reports simulation results for 10^5 rounds of learning on $[+1, -1]$ reinforcement with punishment for the collapsed game corresponding to each strategic class of categorization games we consider. These results show a familiar pattern: a larger bias in state probabilities corresponds to a larger frequency with which the pooling equilibrium is learned. That said, in all treatments of state probabilities, learners were highly successful at avoiding suboptimal equilibria. Even in the worst case, they reached a signaling equilibrium in three quarters of runs and achieved a mean cumulative success rate greater than 0.95.

We simulated 100 runs of 10^5 plays of learning for each possible strategic class of categorization games with four and five atomic states. Results are reported in table 2.

Perhaps unsurprisingly, state probabilities with respect to the reward partition are a signif-

no. atoms	no. available partitions	proportion of partitions in O	strategic class	cum. succ. rate	succ. rate last 1000	$P(\mathcal{S} \in O)$
3	5	0.4	$\{1,2\}$	0.857	0.865	0.725
4	15	0.267	$\{2,2\}$	0.987	0.994	0.977
		0.333	$\{1,3\}$	0.768	0.77	0.41
5	52	0.231	$\{2,3\}$	0.823	0.835	0.644
		0.288	$\{1,4\}$	0.8	0.801	0.293

Table 2.2: Simulation results for categorization games with uniform atomic state probabilities. Results reflect 100 runs of 10^5 plays for each parameter setting.

icant determinant of players' performance in the categorization game. In the $\{2,2\}$ class, whose corresponding collapsed game exhibits uniform state probabilities, cumulative success rates are uniformly close to one. Despite the large number of available categorization schemes, the players learn highly successful signaling and categorization dispositions. By contrast, in the $\{1,3\}$ game, whose collapsed game has biased state probabilities, players are much less successful. The mean success rates (both cumulative and over the final 1000 plays) are close to the expected payoff of 0.75 in the pooling equilibrium of the corresponding collapsed game. Moreover, the sender learns to use a state partition that can support perfect signaling with probability less than 0.5 (but greater than chance performance of 0.33). Similarly, in the five-atomic-state categorization games, the players are more successful in the class whose collapsed game exhibits less biased state probabilities (i.e., the $\{2,3\}$ class).

In the two classes of categorization games whose collapsed games exhibit the most extreme bias in state probabilities ($\{1,3\}$ and $\{1,4\}$), players had the most trouble learning effective categorization schemes. In these games, they learned to use a partition that can support perfect signaling with probability very close to the proportion of such partitions among the total set of available partitions. That is, they did not perform much better than chance at finding and using categorization schemes that make all the practically relevant distinctions

strategic class	proportion of runs ending close to	
	pooling equ.	signaling system
$\{1, 4\}$	0.98	0
$\{1, 3\}$	0.92	0.08
$\{1, 2\}$	0.38	0.58
$\{2, 3\}$	0.36	0.59
$\{2, 2\}$	0	1

Table 2.3: Proportion of runs ending with players' dispositions close to pooling and signaling (separating) equilibria, as determined by cumulative success rate over the final 1000 plays. Runs counted as close to the associated pooling equilibrium if their success rates were within 0.03 of that equilibrium's expected payoff. Runs counted as close to a signaling system if their success rates were greater than 0.95.

in individuating states.

Figure 3 sheds further light on the relationship between category learning and state probabilities, reporting the proportion of runs in which learners in each class of categorization game achieved success rates close to the expected payoffs associated with the pooling and signaling equilibria in the corresponding collapsed game.

There are a few morals suggested by these results. First, as the $\{2, 2\}$ class of categorization games illustrates, merely increasing the number of available categorization schemes need not pose a serious problem for players simultaneously learning to categorize and to signal. Second, the degree of practical success achieved by learners in categorization games with biased probabilities over the reward partition varies systematically with the expected payoff in the pooling equilibrium in the associated collapsed game.

Finally, we should note a worry. In the model above, senders are often successful at finding and using partitions that can support perfect signaling. But they tend to mix over several such partitions, with no particular partition serving as a canonical basis for individuating states. We might say the categorization dispositions they learn are *fuzzy* rather than *sharp*.

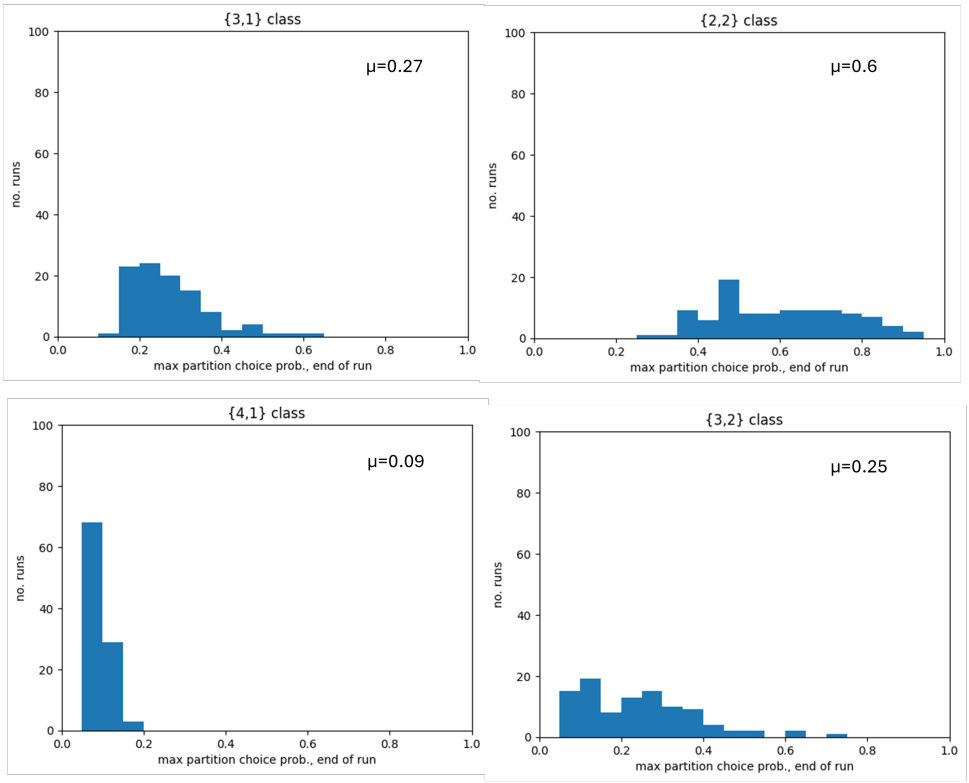

Figure 2.2: Number of runs ending with largest partition probability falling in each 0.05-width bin in the unit interval on reinforcement (with punishment) learning.

One way to see this is by looking at the sender's partition-level choice probabilities at the end of each simulated run. If the largest such probability at the end of a run is close to one, then the sender has learned fairly sharp categorization dispositions—there is just one partition she typically uses. The smaller the largest partition probability, the fuzzier her categorization dispositions. Figure 2 plots the number of runs that ended with the largest partition choice probability falling in each of twenty intervals of length 0.05, for each class of four- and five-atom categorization games. Notice that, for all strategic classes, the sender only rarely learns to use any single partition almost all the time. Partition dispositions are fuzziest in strategic classes with larger biases in their collapsed games' state probabilities, but even in the $\{2, 2\}$ class, the mean of the largest partition probability is just 0.6, with many runs ending with a value below 0.5.

It is reasonable to worry about realism here. Presumably, signalers in nature do not routinely flip-flop among several incompatible categorization strategies in individuating the states they communicate about. We will return to this concern.

2.5 Invention and Memory

One might worry that the learning rule for partition selection bakes in an unrealistic assumption about the cognitive capacities of the sender. In particular, since the sender always has positive propensities for *all* the available partitions, the model seems to assume that the sender starts out with an internal representation of every possible categorization scheme she could use to individuate states, and that she never prunes the set of categorization schemes she considers. This is most worrying in games with larger spaces of atomic states, for which very many partitions exist. Presumably, many learners in nature lack the cognitive resources needed to simultaneously entertain all those partitions as options for individuating states.

To address this concern, we introduce a variant of the model in which partitions are *discovered*, or *invented*, as the sender learns over repeated play of the game, and in which the number of partitions that can be entertained as options at a given time is bounded. The learning dynamics we use is inspired by Alexander et al.'s (2012) *reinforcement with invention*. Rather than modifying a set of propensities defined on an exhaustive set of acts, a reinforcement with invention learner invents the options available to her in the course of learning. In the modified model, we assume that learning at the level of signal and act selection is governed by ordinary reinforcement with punishment learning as in the above model, but that partition learning proceeds by a variant of reinforcement with invention. We also introduce an upper bound on the number of partitions that can be chosen with positive probability at a given timestep. This is meant to reflect a limitation on memory: there are only so many categorization strategies the sender can access as options at a time.

The dynamics works as follows. The partition urn initially contains a single ball, called the *mutator ball*. When the mutator ball is drawn, the sender adds a ball of a new type to her urn, corresponding to an untried partition of the atomic states, and uses the newly-invented partition on that play. For all ball types, each corresponding to a distinct partition of $\mathcal{A}$, reinforcements and punishments work exactly as in the model above (except as modified by the memory constraint discussed below). The mutator ball itself is never reinforced or punished. We assume the order in which partitions are invented is random.

The second component of the modified dynamics is a bound on the number of partitions the sender can simultaneously entertain. We assume that there can be at most k many partitions that get chosen with positive probability at a time. This limitation is enforced in the following way. If, in the course of learning, k partitions get invented and the invention ball is drawn for the $k+1$th time, the learner must *forget* one of the currently active partitions to accommodate the one to be invented. Here, "forgetting" corresponds to assigning probability zero to some partition.

The probability that a given active partition is forgotten varies negatively with the probability that that partition is chosen. The idea is that the sender is less likely to forget a partition the more times she has successfully deployed it in the past. The rule for forgetting works as follows. Suppose there are ten active partitions, each assigned an index between one and ten, with p_i denoting the probability with which partition i is chosen at the present timestep. The mutator ball is drawn, and so one of the active partitions must be forgotten to accommodate the to-be-invented partition. Partition i is forgotten with probability

$$p_{\text{forget}}(i) = \frac{1 - p_i}{\sum_{i=1}^{10}(1 - p_i)}$$

After a partition is forgotten, the new partition to be invented is selected at random from among all inactive partitions, including any that were previously forgotten. As above, we

assume equal reinforcement and punishment values, both set to one.

For the simulations, we set the bound on active partitions to ten. There is nothing special about this particular number. But it is worth noting that, in the contexts we consider, it is a fairly modest constraint. When there are three atomic states, it is compatible with the learner entertaining all categorization options. With four atoms, it allows the sender to consider up to two thirds of the available partitions. Even with five atoms, it allows nearly one fifth of 52 possibilities to be simultaneously contemplated.

Table 4 reports results for learning in the categorization game on the modified dynamics with both invention and limited partition memory. We only consider games with four and five atomic states because the memory limit only makes a difference in games with more than three atoms. Notice that success rates for each strategic class are close to those for the treatment with unlimited memory. Nor are there substantial differences between the two treatments in the final probabilities with which the sender uses a partition that can support perfect signaling. Constraining our players' memory does not seem to substantially hinder their learning to categorize and signal effectively.

no. atoms	no. available partitions	proportion of partitions in O	strategic class	Results: invention with mem. limit		
				cum. succ. rate	succ. rate last 1000	$P(\mathcal{S} \in O)$
4	15	0.267	$\{2,2\}$	0.979	0.988	0.949
		0.333	$\{1,3\}$	0.794	0.795	0.461
5	52	0.231	$\{2,3\}$	0.865	0.874	0.558
		0.288	$\{1,4\}$	0.813	0.813	0.307

Table 2.4: Simulation results for learning in categorization games on reinforcement with punishment learning, with invention and bounded memory for partition selection. Results reflect 100 runs of 10^5 plays for each parameter setting.

The major difference between the two treatments concerns the *sharpness* of the sender's categorization dispositions. Figure 3 is analogous to figure 2, reporting how many runs ended with the largest final partition probability lying in each 0.05-width bin in the unit

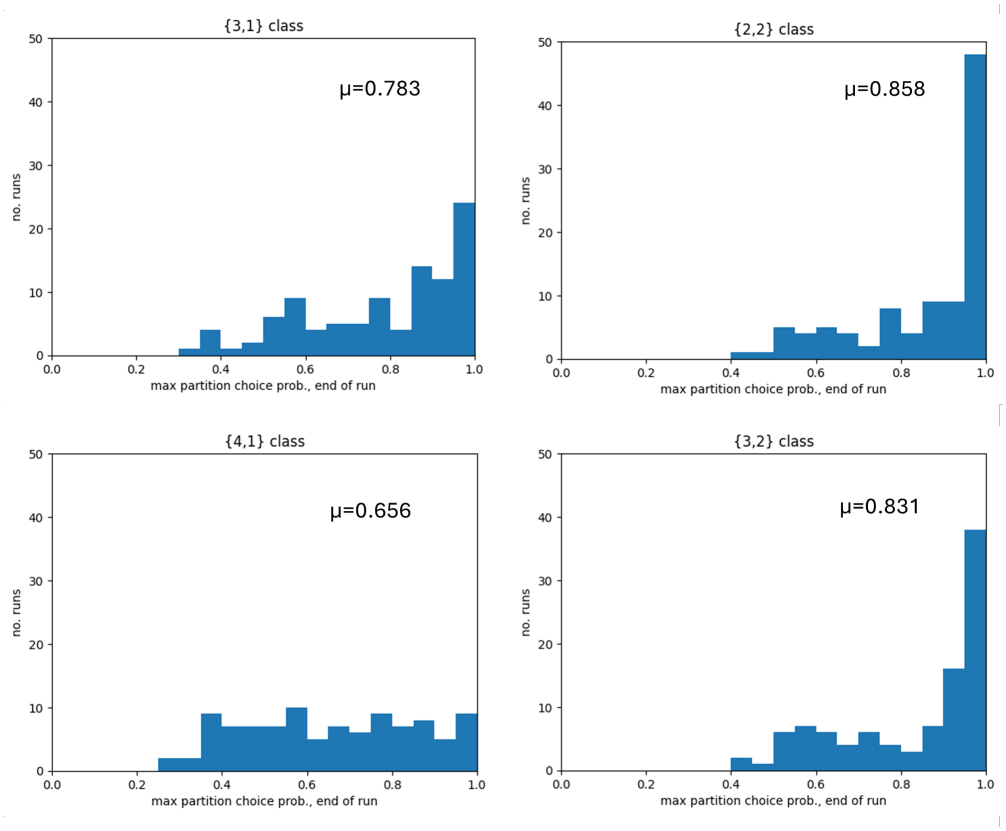

Figure 2.3: Number of runs ending with largest partition probability falling in each 0.05-width bin in the unit interval on reinforcement (with punishment) learning with partition invention and bounded memory. The further to the right a bin is, the sharper the categorization dispositions in the runs it subsumes.

interval, this time for the treatment with bounded memory. Notice that, relative to the model without memory limitations, weight has shifted rightward for all strategic classes. The smallest mean value for runs with limited memory (0.656, for the $\{4, 1\}$ class) is nearly as large as the *largest* such value for the treatment with invention alone (0.677, for the $\{2, 2\}$ class). For $\{2, 2\}$ games, nearly half of runs ended with the sender using a single partition with probability greater than 0.95, and for all classes but $\{4, 1\}$, the modal run lives in the rightmost bin. This reflects significant sharpening relative to the results for treatments without limited memory.

2.6 Discussion and Conclusion

We have considered how the partition of states of nature underlying the sender's conditional dispositions in a signaling game might be learned. We introduced a general framework for modeling category learning in signaling games and considered the results of simulations of several closely-related models developed within that framework.

We find that simple forms of reinforcement learning very often suffice to guide players to adopt effective categorization schemes and signaling conventions. This holds whether we assume the sender starts out entertaining all possible schemes for individuating states, or that she gradually discovers them as she plays.

That said, there a few details worth noting. First, the players' degree of success in learning to categorize states effectively depends on how state probabilities are distributed over elements of the reward partition. In categorization games corresponding to Lewis-Skyrms games with biased state probabilities, the presence of pooling equilibria makes these tasks harder. Second, on the unmodified reinforcement learning dynamics, senders rarely learn very sharp categorization dispositions, insofar as they tend to mix over several candidate partitions. However, when we introduce even a fairly modest constraint on memory with respect to partition selection, players become much more likely to adopt a unique basis for individuating states, at least most of the time.

In Section 2, we highlighted that in many natural settings, the problem of projectibility facing would-be signalers is rendered easier by the guidance provided by evolved perceptual and cognitive biases. The categorization game could be used to model the emergence of categories reflecting these biases by embedding the game in a population model in which selection operates on the partitions used by individual members to individual states. The present focus on learning thus reflects a choice motivated by the goals of the dissertation, not an intrinsic limitation of the modeling framework.

While the lack of help from innate saliences makes the learning problem facing our agents especially unfriendly in one respect, it is worth emphasizing several countervailing assumptions working in the learners' favor. Without attempting an exhaustive catalog: there are never more than five atomic states; there are just two elements in the reward partition; the sender has available exactly two signals; the receiver has available exactly two acts. Relaxing any of these assumptions would yield a more general, but more difficult, setting for category learning in the context of signaling. How reinforcement learners like ours might fare in any of these problems is an open question.

We began by noting the special difficulties that arise in connection with projectibility for signalers. Our results show that, on one way of modeling those difficulties, they can be met without much difficulty by a particularly simple kind of trial-and-error learner.

Chapter 3

Discrimination Learning

3.1 Projectibility for Learners in a Hurry

In the case of the bat scientists, the more effective learner was L., who was attentive to finer distinctions among her observations than was her less successful counterpart, N.[1] But in some contexts, especially when it pays to learn *quickly*, more successful learners are distinguished from less successful learners in the opposite way. Sometimes, attending to too many differentia in one's environment slows down learning, and one could learn better if they only paid attention to one or a few relevant cues.

Consider the following problem. A subject sits in front of a table on which there are two objects, one on the left and one on the right. On each of a series of trials, the left-right position of the objects is randomized, and the subject must choose one or the other. One object is designated as the *winning object*, and the subject receives a reward every time they choose it. When they choose the other object, they get nothing. Every ten trials, the objects

[1]Some material from this chapter is adapted from Torsell and Barrett (2024).

are switched out for a new pair. Problems like this have been a mainstay of experiments on animal learning since the early twentieth century.[2]

There are many ways a learner might approach this problem. We will consider three. Call the first kind of learner an *undiscriminating reinforcement learner*. This learner ignores information about position in choosing an object. They attend only to the quality of the objects. Suppose the objects are a key and a rock. In this case, the learner's choices are determined by propensities defined on just two (unconditional) responses: *choose rock* and *choose key*. They learn by Roth-Erev reinforcement with uniform initial propensities over these two options. We can thus think of their learning in terms of the changing contents of a single urn containing balls labeled *key* and *rock*. When a pair of objects is replaced, their propensities are reset and learning begins anew.

By contrast, a *discriminating reinforcement learner* attends to both position and object quality in selecting an object. They first note the positional layout of the options, and then choose an act conditional on that information. Since they too are a Roth-Erev learner, we can think of their learning in terms of balls and urns. But now we need *two* urns containing *key* and *rock* balls: one the learner draws from when the key is on the left and the rock is on the right, and another they draw from when the objects' positions are flipped. If they succeed on some trial, they only update the contents of the urn corresponding to the positions of the objects on that trial.

The discriminating and undiscriminating reinforcement learners are distinguished by their dispositions to generalize across instances. Just as N. generalizes what he learns from Grey Bat observations to unobserved Indiana Bat instances where L. does not, so does the undiscriminating reinforcement learner generalize what he learns from trials in which the rock is on the left to trials in which the rock is on the right, while the discriminating reinforcement learner does not. Dispositions to transfer learning from one context to another are an

[2]See, e.g., Lashley (1929); Lawrence (1949); Sutherland and Mackintosh (1971).

important determinant of a learner's degree of success in a given problem, especially when speed of learning matters. And, as we will see, learners who generalize on different bases may perform identically in the long run but differ substantially in their short-to-medium run performance.

For the kind of reward-based learners we consider here, there is another aspect of their strategy that matters substantially for speed of learning, which we will call *reward sensitivity.* Reward sensitivity is measured by the magnitude by which a learner's propensities are positively or negatively reinforced upon success or failure in action. The larger the magnitude of these adjustments, the more reward-sensitive the learning strategy. For a highly reward-sensitive learner, a few successful early trials might suffice to crystallize highly determinate dispositions. A less reward-sensitive learner will be more cautious, maintaining an exploratory posture for longer as their dispositions are gradually shaped by incremental adjustments.

Our third kind of learner can be thought of as a maximally reward-sensitive reinforcement learner who, like the undiscriminating reinforcement learner, attends only to object quality in making its selection. In the first trial of a problem, they choose an object at random. If their response is successful, they certainly choose the same object in the next trial. If the action is unsuccessful, they switch to the other option on the next trial. They are *so* reward-sensitive that a single successful or unsuccessful trial locks in their choice in the next trial. This kind of learning has been called *Win-Stay/Lose-Shift* (WSLS).[3]

We can individuate these methods according to the inductive assumptions they reflect. Both the WSLS learner and the undiscriminating reinforcement learner assume that rewards are a function of the type of object selected alone. The discriminating reinforcement learner disagrees, treating reward as a function of *both* object quality and position. We can also

[3]In learning problems with more than two options, things are somewhat subtler. Here, following an unsuccessful trial, a WSLS learner chooses its next response at random from among the set of all available alternatives except the one just used.

gloss the learners' differing reward sensitivities in terms of inductive assumptions. The WSLS learner assumes that if a choice was (not) rewarded on the current trial, then it will certainly (not) be rewarded on the next trial. Both the undiscriminating and the discriminating learner disagree, taking reward outcomes as weaker inductive evidence concerning the underlying reward-choice relationship. Of course, the cognitive language in these descriptions is intended metaphorically: these methods of learning can be implemented by systems without the capacity to explicitly entertain propositions.

In our very simple discrimination learning problem, the pair of inductive assumptions characterizing the WSLS learner turn out to be powerful combination. A WSLS learner will succeed on every trial after the first in a given problem, for an average success rate of 0.95.[4] On simulation, the undiscriminating reinforcement learner succeeds on 0.74 of trials, on average, and the discriminating reinforcement learner enjoys a success rate of 0.65.[5]

While these differences in success rates are significant, they could not be inferred from the long-run behavior of the three learners. Given enough time in a single discrimination problem, all three must eventually learn the optimal response (in the case of the reinforcement learners, "must" means "with probability one").[6] But here they have just ten steps of learning per problem. Speed is crucial. As a consequence, the quick adjustment facilitated by the

[4]Suppose the WSLS agent chooses the rewarded object in trial 1. Then it will choose that object again in trial 2. Suppose it chooses the unrewarded object in trial 1. Then it will pick its trial 2 action from among the set of actions not chosen in trial 1. The only available action is picking the object not picked in the previous trial, i.e. the rewarded object. Since it chooses its first act by a coin flip, it will succeed on nine of ten trials half of the time and on ten of ten trials the other half of the time. See Barrett (2024) for further discussion of WSLS learning in this kind of problem.

[5]Simulations comprised 10^5 problems, each consisting of ten trials of learning.

[6]Footnote 2 suffices to show this is so for WSLS learner. For the two reinforcement learners, it follows from a theorem due to Beggs (2005) that, in the limit of an indefinitely repeated sequence of two-object discrimination problems, both will, with probability one, learn to choose the winning object with probability one. The relevant result: if a Roth-Erev learner confronts a learning problem in which, of its available actions $A_1, A_2, ..., A_n$, there is an option A_i such that, for some constant $\gamma > 1$, the expected reward for choosing A_i always exceeds that of any alternative action by a factor of at least γ, then in the limit, the agent will, with probability one, choose A_i with probability one. In Roth-Erev learning, payoffs are identified with reinforcement magnitudes, so in our problem, the reward for choosing the winning object is always one and the reward for choosing the losing object is always zero. And thus, the *expected* reward for choosing the winning object, whether considered unconditionally or conditional on the positions of the objects, is one, which exceeds the expected payoff of choosing the losing object by *any* factor $\gamma > 1$.

flat-footed assumptions about projectibility embodied in WSLS (that all that matters for choice is the outcome of the previous trial, and that anything worked on the previous trial will work again) is rewarded. The undiscriminating reinforcement learner's more cautious approach is less successful, but it still outperforms the discriminating reinforcement learner, whose sensitivity to reward-irrelevant information slows down learning significantly.

In the 1940s, psychologist Harry Harlow conducted a series of now-famous experiments on learning in two-item discrimination problems like the *key-rock* problem. His subjects were rhesus macaques. He found that, over repeated performance in two-object discrimination problems, the monkeys' intra-problem learning gradually shifted from patterns of gradual improvement, characteristic of some form of reinforcement learning, to the rapid learning associated with WSLS. We might think of this as reflecting the gradual adoption of the two inductive assumptions discussed above in connection with WSLS: the monkeys learn that object quality (but not position) is projectible, and that the outcome of the first trial of a problem should dictate their choices for the rest of that problem. As a result, they learn how to learn more effectively in the kind of problems Harlow gave them.

Here, we consider how such a shift in learning strategy might be achieved. Section 3.2 describes Harlow's project and results. Section 3.3 develops an intermediate model that shows how the monkeys' learning to learn can be understood in terms of the interplay between attentional learning and a gradual ratcheting-up of reward sensitivity. Section 3.4 shows how such a shift could be achieved through a form of hierarchical reinforcement learning, where reward sensitivity itself is shaped by practical success and failure.

3.2 Harlow's Experiments

Harlow (1949) discusses three experiments involving a group of eight rhesus monkeys. Harlow

was interested in the formation of *learning sets*:

> The learning of primary importance to the primates, at least, is the formation of learning sets; it is the learning how to learn efficiently in the situations the animal frequently encounters. This learning to learn transforms the organism from a creature that adapts to a changing environment by trial and error to one that adapts by seeming hypothesis and insight. (Harlow, 1949, 51)

Harlow's goal was to "demonstrate the extremely orderly and quantifiable nature of the development of certain learning sets and, more broadly, to indicate the importance of learning sets to the development of intellectual organization and personality structure" (Harlow, 1949, 52). As we will see, he succeeded in dramatic fashion.

Here, we are concerned with Harlow's first experiment, which presented his subjects with a series of object quality discrimination (OQD) problems. On each trial of an OQD problem, a monkey sat in a testing apparatus in which it was presented a tray containing two objects, one covering a right-hand food well, the other covering a left-hand food well. The monkey then removed one of these objects and received the contents of the food well beneath. One of the wells held a food reward; the other contained nothing.

A single pair of objects was used for every trial in a given problem (different object pairs were used in different problems), and which object covered the reward was held constant across trials. The spatial position of that object, however, was randomized between left and right. Harlow presented each monkey with 344 problems of this kind. Problems 1-32 consisted of 50 trials each, 33-234 were run for just six trials, and problems 235-344 had an average length of nine trials.

Average success rates for early trials of blocks of problems in the first experiment are reported in figure 1. In early problems, the monkeys' average success grows gradually between trials one and six. But as they receive more training, the curve steepens dramatically. By the final

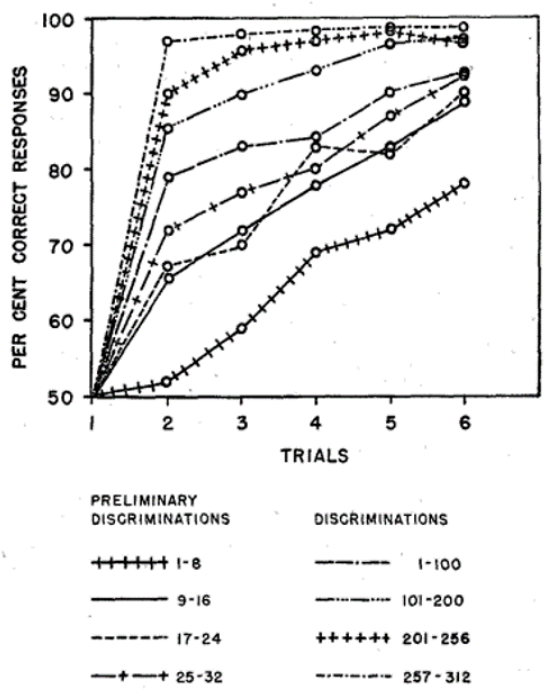

FIG. 2. Discrimination learning curves on successive blocks of problems.

Figure 3.1: Average success for early trials of experiment 1 (reprinted from Harlow (1949))

block of problems (257-433), the monkeys chose nearly optimally by the second trial. Harlow says of the evolution of the monkeys' learning strategies as the development of a learning set for two-item OQD problems: "The monkeys learn how to learn individual problems with a minimum of errors. It is this learning how to learn a kind of problem that we designate by the term learning set" (Harlow, 1949, 53).

Developing the right attentional saliences is essential to success in this kind of problem. The monkeys must learn to attend to object quality, rather than position, in order to reliably succeed in this kind of problem. There is evidence that location is naturally salient for some organisms (see, e.g., (Zentall and Riley, 2000, 51)), so learning to attend to object quality may involve the nontrivial task of overcoming a default disposition to attend to object position.

3.3 From Reinforcement Learning to Win-Stay/Lose-Shift

Recall that early-trial success rates in Harlow's first experiment showed a transition from smooth, gradual learning to faster "trial-and-error" learning. Barrett (2025) describes this process as approximating a shift from reinforcement learning to Win-Stay-Lose-Shift (WSLS).

As was illustrated by the rock/key example above, in a learning problem with just two available actions, one of which is always rewarded and one of which is never rewarded, a WSLS learner will always succeed in every trial after the first (see footnote 2). In the first model of OQD problem, this means a WSLS learner that has learned to always attend to object quality will certainly be rewarded in every trial after the first. Late-problem learning in Harlow's first experiment comes close to this benchmark: Figure 3.1 shows that the average success rate in trial two for problems 289-344 is approximately 0.97, and it gradually increases to a value in the neighborhood of 0.99 by trial six.

Harlow described the gradually steepening learning curves as illustrating the animals' "deliver[ance]...from Thorndikian bondage" (Harlow, 1949, 59). One moral from the models presented below is that this description is potentially misleading. Basically Thorndikean machinery suffices to replicate the learning to learn achieved by Harlow's monkeys, as long as we treat the sensitivity of the learner's dispositions to rewards and punishments as itself subject to learning.

Commenting on Harlow's results, Barrett (2025) suggests a process by which a reinforcement learner might transform into a WSLS learner. Suppose a reinforcement-with-punishment learner faces a repeated problem in which she chooses between two actions. Suppose action weights are unbounded above and below (i.e., in an urn model, if ball types corresponding to each action are allowed both to run to extinction and to grow arbitrarily large). Allowing the values of both the reinforcement and punishment parameters to increase to arbitrarily large values produces learning behavior functionally equivalent to WSLS in the limit. A

dynamics on which an agent's learning algorithm *itself* evolves according to some rule that ratchets up reinforcements and punishments over time could generate a transformation from simple reinforcement learning (or something close to it) to WSLS-like learning.

Here, we develop a model implements Barrett's suggestion in a context in which the learner's saliences *coevolve* with its learning algorithm. This model considers how an agent might learn to learn better across problems, as Harlow's monkeys did. As in Harlow's experiment, in the model an individual problem is a sequence of OQD trials with a unique pair of objects.

We first describe a model of learning within problems, and then consider how to model the gradual increase in reinforcement and punishment levels across problems. The monkeys' learning is modeled as a two-level process in which both attentional dispositions and first-order choice dispositions evolve by reinforcement with punishment.

Let $A = a_1, a_2, ..., a_n$ denote the actions available to an agent facing some learning problem, and let $q : A \to \mathbb{R}$ be the function mapping each action to its weight. Use q_i to abbreviate $q(a_i)$. Each a_i is chosen with probability

$$p_i = \frac{q_i}{\sum_{j=1}^{n} q_j}$$

Let $q_i(t)$ denote a_i's weight at time t. Suppose that payoffs are bivalent, represented by a function u such that $u(t) = 1$ if the learner's choice at t was successful and $u(t) = 0$ otherwise. The update rule specifies how action weights grow in response success and shrink in response to failure:

$$q_i(t+1) = \begin{cases} q_i(t) & \text{if } a_i \text{ was not chosen at } t \\ q_i(t) + i & \text{if } a_i \text{ was chosen at } t \text{ and } u(t) = 1 \\ q_i(t) - j & \text{if } a_i \text{ was chosen at } t \text{ and } u(t) = 0 \end{cases}$$

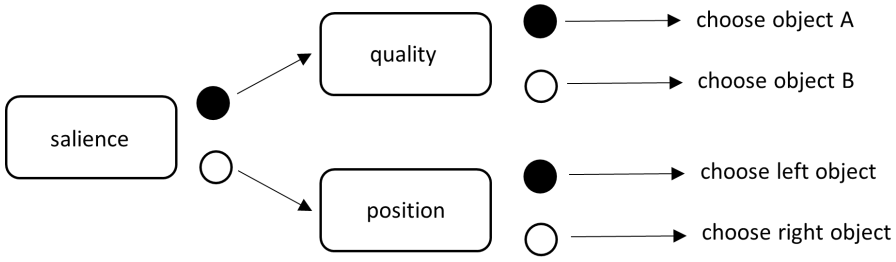

Figure 3.2: The urn model

i is the reinforcement parameter and j is the punishment parameter, and we say an agent learns by $(+i, -j)$ reinforcement with punishment. If $i = 1$ and $j = 0$, we say an agent is implementing *simple reinforcement learning* (see, e.g., Roth and Erev (1995)). An agent's learning is fully specified if, in addition to assigning determinate values for i and j, we specify initial weights, i.e., values for $q_i(0)$ for all $1 \leq i \leq n$.

On every trial of a problem, the position of the objects is randomized, and the reward is always placed under the pre-determined winning object, in whichever food well it happens to cover. The learner chooses a response in the following way. First, it draws a ball from a *salience urn* containing white balls and black balls. This determines which stimulus dimension the monkey attends to in determining its action. For simplicity, we assume that on any given trial, the monkey has just two attentional options: attend to quality or attend to position. Initially, the urn contains one ball of each color.

If the learner pulls a black ball from the salience urn, it determines which object to pick up by a random and unbiased draw from another urn, called the *quality act urn.* The quality act urn also contains black and white balls, initially holding one of each. If the monkey pulls a black ball, then it picks up object A; if it draws a white ball, it picks up object B. If the object the monkey picks up is the reward object, then for both the salience and quality act urn, the monkey returns the drawn ball and adds i many new balls of its type (i is the reinforcement parameter). If it picks up the non-reward object, it returns the drawn ball and then removes j many balls of its type, for both the salience urn and the quality

act urn, provided that removing j many balls would not drive that type to extinction. If punishment by removing j many balls *would* drive a type to extinction, then balls of that type are removed until exactly one remains in the relevant urn. Weights for each attentional and choosing disposition are thus bounded below by one, preventing possible strategies from being completely eliminated.

The process is analogous if a white ball is drawn from the salience urn. In that case, the monkey selects which object to pick up by a random and unbiased draw from a third urn, the *position act urn*, with the same initial contents as the quality act urn. If a black ball is drawn, the monkey picks up the object on the left; if it pulls a white ball, it picks up the object on the right. Reinforcement (both positive and negative) operates exactly as in the first case.

Notice that the monkey's first-order object choices (and thus the character of the dispositions reinforced and punished) depend on its attentional saliences: if it attends to quality, its choice is between picking up object A and picking up object B; if it attends to position, it chooses between picking up the object on the left and picking up the object on the right.

Transformation of reinforcement and punishment parameters is accomplished in the following way. Let i_n and j_n denote, respectively, the reinforcement and punishment parameters characterizing the agent's learning in the nth problem in a sequence of OQD problems. Let $\alpha > 1$ be a constant. The agent's learning algorithm evolves according to the following rule:

$$i_{n+1} = \alpha i_n$$

$$j_{n+1} = \alpha j_n$$

The iterated rescaling used here is just one way of realizing the desired growth of reinforcement and punishment parameters over time. The qualitative interpretation is that the growing parameters reflect the monkeys' improving grasp of the structure of the problem

they're facing—that this is a problem in which choosing one kind of object *always* leads to success (so no need to try out other strategies after you've picked the right object) and choosing another kind of object *never* does (so you shouldn't pick up an initially-unrewarded object after you've chosen it once).

Simulations were run with two different treatments of these iterated transformations. In the first variant, the algorithms governing both salience and act-level learning are set to the same initial parameters and evolve together according to the rule above. In the second, saliences and act-level dispositions evolve according to the same dynamics in problem 1 (i.e., reinforcement and punishment parameters are set to i_1 and j_1, respectively), but in subsequent problems their values are transformed *only for act-level learning* (i.e, saliences always evolve according to $(+i_1, -j_i)$ reinforcement with punishment). These variants correspond to subtly different coevolutionary stories. In the first, the agent's learning approaches WSLS with respect to both attention and action. In the second, attentional learning is still transferred across problems, but it always evolves according to its original more gradual learning rule while act-level learning shifts toward WSLS.

There are two other assumptions to note. First, the values for both reinforcement and punishment evolve in lockstep, undergoing the same transformation at the conclusion of each problem. Second, both attentional and act-level dispositions evolve according to the same dynamics, at least initially (in the first variant, this is true only in problem 1; in the second variant, balls are added and removed according to the same rules for all three urns in every problem). These are both simplifying assumptions; nothing essential to the model depends on them.

At the conclusion of each problem, the contents of both act urns are returned to their original state (one black ball and one white ball). This is meant to reflect the monkeys' recognition that the appearance of a new pair of objects signals the start of a *new problem*—their training on the previous problem gives no clue as to which will be rewarded. The contents

of the salience urn are not reset. This is meant to capture the transfer of salience learning across problems, as documented in Harlow's experiments. Weights for all acts (including "attentional acts") are unbounded above and bounded below by 10^{-14}.

Simulations of 1000 runs were executed on various parametrizations of the model. An individual run comprises 344 problems. The first 32 problems consist of 50 trials each, followed by 200 six-trial problems and 112 nine-trial problems, following Harlow's setup in experiment one. To assess the ability of the model to capture Harlow's results in that experiment, special attention was paid to average success rates in the first six trials of each problem.

For both variants of the model, the following parametrization achieved a close fit with Harlow's data:

$$i_1 = 1$$

$$j_i = 0.25$$

$$\alpha = 1.02$$

For both variants, these parameters generate a sequence of learning curves following the steepening pattern observed in Harlow's first experiment. The qualitative fit between the data generated by simulated runs of the first variant and Harlow's experimental data is illustrated in figure 4. Figure 3.3, which reports mean (per-trial) absolute difference in mean success rates for trials 2-6 between Harlow's data and simulated runs of both model 2 variants, quantifies the closeness of the fit. For no block of problems does this difference exceed 0.06, and for most blocks, it is considerably smaller than that. The model, on both treatments, very closely approximates Harlow's monkeys' learning. The closeness of the fit indicates that the model not only shows how a reinforcement learner might in principle learn

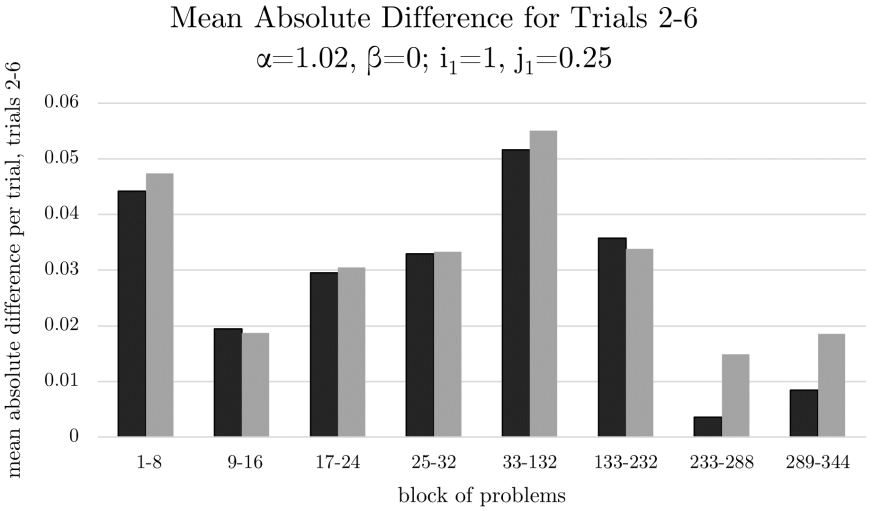

Figure 3.3: Mean absolute error for trials 2-6. Results for the first variant (both salience and act learning transformed) are in black; results for the second variant (only act learning transformed) are in grey.

to implement WSLS in the context of a certain kind of problem; it is capable of mimicking the behavior of flesh-and-blood subjects in a classic series of experiments concerning learning to learn.

It should be noted that it is not possible to capture the monkeys' late-problem learning without positive punishment. In the final two blocks of problems, Harlow's monkeys choose correctly over 90 percent of the time by trial 2. While it is true that expected trial 2 success rate (conditional on attending to the reward-relevant dimension) varies positively with the ratio of reinforcement level (i) to the initial number of balls in the quality act urn, reinforcement alone is not enough to get learning as fast as Harlow's monkeys achieve in the final blocks of experiment one. For a reinforcement (without punishment) learner with optimal saliences, the expected success rate in trial two is bounded from above at 0.75.[7]

[7]Suppose that with probability 1 a reinforcement learner attends to the relevant dimension in every trial of problem n. Set initial contents for act urns as above. The agent will choose the rewarded object in trial 1 with probability 0.5. Let i_n be very large, so that the probability that the agent chooses the rewarded object in trial 2 conditional on having chosen the correct object in trial 1 is 1-ϵ. If $j_n = 0$, then the agent will choose the rewarded object in trial 2 with probability 0.5, conditional on having chosen the unrewarded object in trial 1. Thus the probability of success in trial 2 is given by $0.5 \cdot (1 - \epsilon) + 0.5 \cdot 0.5$. Letting ϵ run to zero, this value asymptotically approaches 0.75 from below.

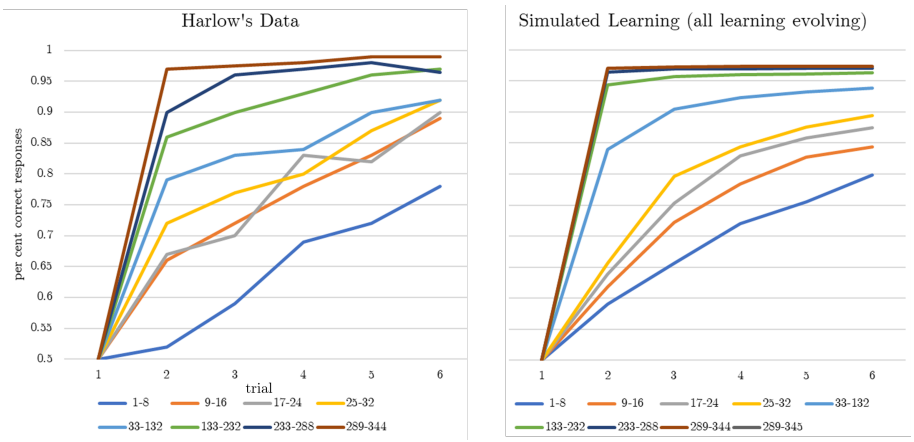

Figure 3.4: Harlow's data compared with learning curves generated by model 2 (variant 1) simulation with $i_1 = 1$, $j_1 = 0.25$, $\alpha = 1.02$

Without punishment, no level of reinforcement, no matter how large, can generate a success rate of 0.97 in trial 2, as observed in the final problems of Harlow's first experiment.

It is also worth noting the chief difference between the two variants' performance: late-problem learning is less success in the second variant (only act learning rescaled) than in the first (salience and act learning coevolve). Compare the learning curves for problems 289-344 in the two graphs of figure 5. Both runs were executed with $\alpha = 1.02$ and initial reinforcement and punishment values of 1 and 0.25, respectively, for both salience and act-level learning. While learning curves for early and middle blocks are nearly identical, there is a clear—albeit small—-divergence between learning in the final blocks across the two variants.

The reason is that optimal saliences become more firmly cemented when reinforcement and punishment levels for both attention-level and act-level learning grow together. In 10^4 runs of the second model ($\alpha = 1.02$, $i_1 = 1$, $j_1 = 0.25$, as above) with transformations applying to act-level learning only, the composition of the salience urn at trial six of problem 289 (the first problem of the final block for which Harlow reports success rates) was biased approximately 20.5:1 in favor of attending to object quality, on average, increasing to about 22.5:1 by trial six of problem 344. Applying transformations to salience learning as well produces average object-quality biases of approximately 40.5:1 and 55:1 for the same problems. The faster

final-block learning in the first variant more closely approximates Harlow's data than the second variant's slower learning.

Notice that this is not a very significant difference. That the two variants perform so similarly suggests that the primary driver of the modeled monkeys' learning to learn lies in the transformation of act-level learning, rather than salience learning, across iterated problems.

3.4 Learning to Learn Better

Our model closely approximates the learning to learn achieved by Harlow's monkeys. There is good reason, however, to hesitate in taking the model as illustrating *learning* to learn, in Harlow's sense. The modeled agent's salience- and act-level learning is transformed by iterated rescaling of reinforcement and punishment values, but those transformations occur automatically between problems, independently of the agent's experience. In what we might call a *genuine learning process*, one should expect how an agent's dispositions change over time to be sensitive to what she experiences. And this experience-sensitivity should be non-trivial in the sense that the evolution of the agent's dispositions, in this case transformations of the agent's first-order learning, is not determined by the time-step of the learning process alone, as in the model above.

Ideally, we would like to model the genuine learning process *actually* underlying the evolution of the individual monkeys' first-order learning in Harlow's experiment. But to construct such a model, we would need data concerning each monkey's individual performance in the context of a variety of learning problems. Unfortunately, Harlow reports only aggregated data from his eight subjects for a few discrimination problems. That said, we can model a plausible learning process that *would* lead an agent to gradually shift from implementing slower reinforcement learning to fast win-stay/lose-shift-like learning in Harlow-style discrimination

problems.

Consider an agent facing series of n many k-trial object quality discrimination problems, structured just as in Harlow's experiment. In the new model, the agent's choice behavior and first-order learning proceeds exactly as in our first model: her saliences and act-level dispositions evolve by reinforcement with punishment, as captured by the urn model in figure 3.2. The new model diverges from the old in how the updating of the first-order learning parameters is handled. In the new model, the agent is equipped with a higher-order learning mechanism that modifies its first-order reinforcement and punishment levels differently depending on its choices and outcomes in each pair of contiguous trials. One might think of how first-order reinforcements and punishments are updated as a form of higher-order reinforcement learning.

The basic idea behind the mechanism is simple but requires some setup. In order to describe how higher-order learning works, we will say that a first-order action is *successful* if the agent chose it in the immediately preceding trial and it led to success, and we will say that an action is *unsuccessful* if the agent chose it in the immediately preceding trial and it led to failure.

Note that for any fixed initial assignment of propensities over first-order acts, the higher a reinforcement learner's level of reinforcement, the *more* likely she is to repeat successful actions; and the lower her level of reinforcement, the *less* likely she is to repeat successful actions. Thus, increasing the level of reinforcement can be thought of as reinforcing the disposition to repeat actions which just led to success, and decreasing the level of reinforcement can be thought of as punishing the disposition to repeat actions which just led to success. Similarly, for any fixed initial assignment of propensities over first-order acts, the lower the learner's level of punishment, the *more* likely she is to repeat unsuccessful actions; and the higher the level of punishment, the *less* likely she is to repeat unsuccessful actions. So, decreasing the level of punishment can be thought of as reinforcing the disposition to

repeat actions which just led to failure, and increasing the level of punishment can be thought of as punishing the disposition to repeat actions which just led to failure.

The higher-order dynamics that describes the evolution of the first-order learning parameters is a variety of reinforcement learning in which the agent reinforces and punishes four different higher-order learning strategies: (i) stick with strategies when they succeed, (ii) abandon strategies when they succeed, (iii) stick with strategies when they fail, and (iv) abandon strategies when they fail.

The higher-order learning dynamics is as follows: the agent increases her level of reinforcement whenever she succeeds upon repeating a strategy that just led to success; she decreases her level of reinforcement whenever she fails upon repeating a strategy that just led to success; she decreases her level of punishment whenever she succeeds upon repeating a strategy that just led to failure; and she increases her level of punishment whenever she fails upon repeating a strategy that just led to failure.

To formally specify the higher-order learning dynamics, it will be convenient to introduce some new notation. As above, let i_t stand for the level of reinforcement for salience and act learning at time t, and let j_t represent the corresponding level of punishment at t, where timesteps are cumulative across problems (so that, e.g., the first trial of the second problem occurs at $t = k + 1$, not $t = 1$). Let $s(t)$ denote the salient dimension on that trial, i.e. the feature of the stimulus objects the learner attends to at t, and let $I_s(t, t + 1)$ be a function whose value is 1 if $s(t) = s(t + 1)$ and 0 otherwise. Let $o(t)$ denote the outcome of the trial at timestep t, where $o(t) = 0$ if the trial was unsuccessful and $o(t) = 1$ if the trial was successful. $a(t)$ will denote the act chosen at t, and $I_a(t, t + 1)$ is a function whose value is 1 if $a(t) = a(t + 1)$ and 0 otherwise. $\gamma > 1$ and $\lambda < 1$ are constants. Figure 3.5 specifies how reinforcement and punishment values change as a function of I_s, I_a, and o on the metalearning dynamics.

$I_s(t,t+1)$	$I_a(t,t+1)$	$\langle o(t), o(t+1)\rangle$	i_{t+1}	j_{t+1}	qualitative description
1	1	$\langle 0,0\rangle$	i_t	γj_t	act chosen at t repeated at $t+1$, both trials unsuccessful; learner reinforces disposition to switch from unsuccessful actions
1	0	$\langle 0,0\rangle$	i_t	λj_t	act chosen at t not repeated at $t+1$, both trials unsuccessful; learner punishes disposition to switch from unsuccessful actions
1	1	$\langle 0,1\rangle$	i_t	λj_t	act chosen at t repeated at $t+1$, first trial unsuccessful but second successful; learner punishes disposition to switch from unsuccessful actions
1	0	$\langle 0,1\rangle$	i_t	γj_t	act chosen at t not repeated at $t+1$, first trial unsuccessful but second successful; learner reinforces disposition to switch from unsuccessful actions
1	1	$\langle 1,0\rangle$	λi_t	j_t	act chosen at t repeated at $t+1$, first trial successful but second unsuccessful; learner punishes disposition to stick with successful actions
1	0	$\langle 1,0\rangle$	γi_t	j_t	act chosen at t not repeated at $t+1$, first trial successful but second unsuccessful; learner reinforces disposition to stick with successful actions
1	1	$\langle 1,1\rangle$	γi_t	j_t	act chosen at t repeated at $t+1$, led to success both times; learner reinforces disposition to stick with successful actions
1	0	$\langle 1,1\rangle$	λi_t	j_t	act chosen at t not repeated at $t+1$, both trials successful; learner punishes disposition to stick with successful actions
0	—	—	i_t	j_t	shift in agent's framing of choice situation between t and $t+1$; no update

Table 3.1: The higher-order learning dynamics.

Unlike the earlier model, reinforcement and punishment levels might get modified on any given trial, though there is no case in which both levels are modified on the *same* trial. In the first model, reinforcement and punishment levels only change between *problems*, and they always shift together in lockstep.

Notice that in the new model, reinforcement and punishment levels remain unchanged whenever the agent's saliences shift between contiguous trials (i.e., whenever $I_s(t,t+1) = 0$). The idea is that which feature is salient to the agent in a given trial determines a *framing* of the choice problem at hand, and choices made in trials with different framings are not comparable. In particular, it is not meaningful to treat the sequence of choices made in two contiguous trials in which the learner switched saliences as an instance of *staying with* or *shifting from* a given strategy.

An example may be helpful. Consider a Harlow discrimination problem in which the two stimulus objects are a red bowl and a green cup. In trial t, object quality is salient to the learner. It therefore frames the problem at hand in terms of choosing the right kind of object. On trial $t+1$ the learner's salience changes. Now, it attends to position, and thus sees the problem as requiring it to choose the correct *position in space*.

The example is meant to illustrate that a shift in salience marks a change in how the learner frames its options: in trial t, it faces a choice between different types of objects; in trial $t+1$ it faces a choice between different spatial locations. Suppose it first (at t) chooses the green cup, and then (at $t+1$) chooses the right-hand position, which happens to be occupied by the green cup. Of course, from an outside observer's perspective, the same object was selected between the trials. But from the learner's perspective, the acts *choose the right-hand position* and *choose the green, cup-shaped object* are not comparable such that it makes sense to speak of the learner *shifting* or *staying* between t and $t+1$ to be meaningful. It therefore makes little sense to suppose that the agent updates its dispositions to stay with or switch away from successful or unsuccessful actions after two-trial sequences in which the agent's saliences change between trials.

To investigate whether the new model can can accomplish the desired shift from gradual reinforcement learning to faster win-stay/lose-shift-type learning, we ran a series of 1000 simulations, each consisting in 1000 blocks of 10-trial Harlow discrimination problems. Propensities for acts and saliences were bounded from below at 0.01. All initial propensities were set to 1. γ was set to 1.02 and λ was set to 0.98. Initial reinforcement and punishment levels were $i_1 = 1$ and $j_1 = 0.25$, as in the earlier model.

Results show that the modeled agent reliably learned to use large reinforcement and punishment levels allowing it to approximate win-stay/lose-shift. The mean cumulative success rate over all 1000 problems, averaged across the 100 runs, was 0.91. Restricted to the last 100 problems, the mean success rate across all runs rises to 0.929. This is close to the

optimal expected success rate of 0.95 for a true win-stay/lose-shift learner in this problem: recall that (conditional on its attending to the task-appropriate dimension) such a learner will succeed on the first trial of a given problem half the time on average, and will succeed on every subsequent trial in that problem.

The mean cumulative success rate on the last hundred problems was less than 0.9 on just 47 of the 1000 runs. For all but two of these runs, this value was between 0.47 and 0.52.[8] Why had the agent learned to perform no better than chance on these runs? The explanation has to do with salience. We tracked the proportion of "object quality" balls in the learner's salience urn at the end of each run. This gives the final probability with which the agent chose to attend to the task-relevant dimension. For all 45 of the runs in which the learner performed very close to chance in the last hundred problems, this value was very close to zero (all values ≤ 0.007).[9] This means that on these runs, the learner had glommed onto the wrong salience for the problem at hand.

When the 47 outlier runs are removed, the mean success rate over the final hundred problems of each run rises to 0.954. This indicates that in most runs, the agent successfully glommed onto the task-appropriate salience (*attend to object quality*) and learned to closely approximate Win-Stay/Lose-Shift.

3.5 Conclusion

We began by considering simple discrimination problems with a short window for learning, and we showed how a learner's degree of success in such problems depends on the inductive assumptions embodied in their method of learning. Then, we considered a empirical case in which monkeys confronting a series of such problems appear to gradually adopt a learning

[8] For the two outliers, success rates for the last hundred problems were 0.57 and 0.62.

[9] In one of the two outlier runs, the agent had learned to attend to object quality with probability very close to 1; in the other, the agent's final probability of attending to object quality was 0.377.

strategy that reflects assumptions well-suited to these problems, namely, Win-Stay/Lose-Shift. As in the previous chapter, we showed how a very simple form trial-and-error learning might suffice to solve the problem of projectibility at hand—in this case, learning to approximate Win-Stay/Lose-Shift. If it is appropriate to describe our modeled learner as being "deliver[ed] from Thorndikean bondage," as in Harlow's gloss, then we should add the qualification that it was delivered by Thordikean means.

Chapter 4

Task Switching

4.1 Individuating Choice Problems

Many models of learning assume an agent facing a sequence of identical trials of a well-defined decision problem. But the practical lives of humans and other natural learners are rarely, if ever, so simple.[1] When we learn in the course of repeated experience in a certain kind of problem, the individual instances we encounter are embedded in a varied stream of choice situations presented by our environment. And in navigating these situations, nature does not specify a unique basis on which to make judgments of similarity between them. No two choice settings will be identical in every respect, and any two choice settings will be analogous to each other in some respects. We must *learn* which observable similarities between different choice settings are indicative of a common underlying structure, such that treating those settings as identical for the purposes of learning and action is conducive to practical success.

[1] Much of the material in this chapter is adapted from Torsell (2026).

This is a problem of projectibility. Experiments on *task switching* offer vivid examples of human and animal learners confronting a version of this problem.[2] In task switching problems, a subject is initially trained to perform two different tasks individually. Then, they face a series of trials in which they are sometimes required to perform the first task and sometimes required to perform the second task. A subject can reliably succeed in these problems only if she learns to flexibly toggle between distinct sets of dispositions, each appropriate to one task, depending on the state of a designated task cue. This, in turn, requires that the subject notice and condition her action on the task cue, rather than ignore it or attend to some other, uninformative feature of her environment to decide which task to execute.

Here, we model how an agent might learn both to execute and distinguish between tasks in a simulated training environment based on experiments on task switching in rhesus macaques reported in Avdagic et al. (2014). In the model, learning is characterized in terms of the interactions among three subagents who are responsible for determining the agent's attentional and behavioral responses. The subagents' interests are aligned: each gets a positive payoff only when the learning system comprising all of them successfully performs the relevant task. The subagents' dispositions evolve by a simple form of reinforcement learning.

The goal is to show how the problem of projectibility illustrated by task switching problems might be solved by low-rationality means widely available to natural learners. We do not, however, claim to be describing how the monkeys in fact learned in Avdagic et al.'s experiments. They may well have made use of more sophisticated (or, at any rate, qualitatively different) cognitive machinery than we allow in the model. What matters for our purposes is that the simple, naturalistically motivated dynamics we consider is *enough* to learn to manage the special demands of task switching problems.

As the modeled agent's attention gradually comes to be focused on the practically relevant

[2]See Monsell (2003); Kiesel et al. (2010) for an overview.

regularities in her environment, she becomes a more effective learner in the problems she repeatedly faces. The phenomenon modeled here is thus another instance—alongside our example from Harlow in the previous chapter—of learning how to learn better in a particular kind of problem. Moreover, the basic mechanism behind the metalearning is a form of higher-order reinforcement learning, in which reinforcements at the attentional level shape the agent's dispositions to choose different task-specific behavioral strategies. Models like this, in which reinforcements accrue to full behavioral strategies rather than individual acts, are not without precedent. Erev and Barron's (2005) RELACS (REinforcement Learning Across Cognitive Strategies) is a well-known example. In their model, agents choose from among three decision rules to guide their first-order choices in an iterated binary decision problem, with their propensities over decision rules evolving by reinforcement learning. In a related project, Stahl (1996), building on Rosenthal's (1993) work on "rules of thumb" in games, considers reinforcement learning over strategic heuristics in a p-beauty contest, where the rules under consideration derive from a model of boundedly rational choice developed in Nagel (1995). In both Stahl's and Erev and Barron's models, the first-order strategies over which the agents learn are fully specified in advance by the modeler. By contrast, in our model, the task-specific behavioral strategies themselves emerge from a reinforcement learning process.

4.2 More Monkey Learning

Avdagic et al. (2014) investigate task switching in rhesus macacques facing a series of irregularly shifting *simultaneous chaining* (SimChain) tasks. On each trial of a SimChain task, a subject is presented with a collection of stimulus items which vary along some discriminable dimension (e.g., size or color). If the subject successively selects (e.g., by touching or pointing) the individual items in an order reflecting their ranking with respect to the designated

dimension, then it receives a reward. The reward condition may require the subject to select the items in either ascending or descending order. A SimChain task may, for example, require subjects to successively touch individual colored cards in light-to-dark order in order to receive a food reward. Subjects are only rewarded upon successfully ordering the entire collection of items. No rewards are received in trials in which the subject selects some but not all items in the task-appropriate order.

Avdagic et al.'s experiments presented three rhesus monkeys—named Lashley, Oberon, and MacDuff—with a series of trials which shifted randomly between two SimChain tasks. In both tasks, a subject was presented simultaneously with k many stimulus items on a screen ($3 \leq k \leq 6$). The items were randomly scattered spatially, and each consisted of a white rectangle with a grey circle in the middle. The circles varied with respect to two discriminable dimensions: luminosity (ranging from light grey to black) and radius. In the first task, subjects were rewarded for successively touching the stimuli in an order reflecting their luminosity ranking. Call this the *luminosity task*. In the second task, subjects were rewarded for ranking the stimuli according to radius. Call this the *radius task*.

With respect to the luminosity task, we call luminosity the *target dimension* and radius the *distractor dimension*. These designations are flipped for the radius task. In both tasks, Lashley and Oberon were rewarded for making selections reflecting increasing order (darkest-to-brightest in the luminosity task; smallest-to-largest in the radius task), while MacDuff was rewarded for selections reflecting decreasing order (brightest-to-darkest or largest-to-smallest).

The *training stage* of Avdagic et al.'s experiment consisted of three phases. In each phase, stimulus items were always presented on a *blue* background for the luminosity task trials and on a *red* background for the radius task trials. In the first phase, each monkey faced only radius task trials, and the task was simplified by allowing stimulus items to vary only with respect to radius (the reward-relevant dimension)—luminosity was held constant across

items. If a subject correctly ranked the items according to the radius, it received a food reward immediately after the final item was selected. If it selected an item out of order, the trial was ended immediately and the subject experienced a timeout.

Each subject first faced repeated sessions of 50 three-item trials until it reached an 80 percent performance criterion (i.e., until it succeeded on 40 of the 50 trials in a session), then progressed to sessions of four-item trials, working up to six-item trials with the number of items increasing by one each time the subject reached the 80 percent criterion. This phase ended when each subject reached a 50 percent performance criterion on six-item trials.[3] The second training phase proceeded exactly like the first but with the luminosity task as the target task; here, stimuli varied with respect to luminosity but not radius.

In the third training phase, trials randomly switched between the luminosity task and the radius task, each task being chosen with equal probability on every trial. As in the first and second phases, the stimulus items presented varied only with respect to the reward-relevant dimension for a given trial. The monkeys had to learn to flexibly switch between the patterns of behavior they had learned for the luminosity task and the radius task to consistently receive rewards. They could achieve this either by conditioning their choices on the color of the background or on which feature varied across objects. Again, sessions of 50 trials were repeated until an 80 percent performance criterion was reached. But in this phase there was no progression from smaller to larger collections of stimulus items. Instead, the number of stimulus items varied between three and six from trial to trial in every session.

The training stage was followed by the *experimental stage*, in which stimulus items varied in both dimensions in every trial. The experimental stage consisted of two phases. In the first phase (lasting 36 sessions, or 1800 trials), subjects faced *biased* versions of the luminosity task and the radius task, in which the differences between stimulus items with respect to the distractor dimension were reduced relative to differences in the target dimension. This was

[3] Avdagic et al. do not explain why the performance criterion changed for the final segment of this phase.

meant to increase the salience of the target dimension. Trials were randomly and uniformly distributed with respect to task type (A or B) and number of items to be ranked (three or four). The second phase (lasting 48 sessions, or 2400 trials) proceeded exactly like the first but without bias in the magnitudes of the differences among stimuli along the two dimensions.

All three of Avdagic et al.'s subjects successfully completed all phases of the training stage. However, the authors do not provide data on the monkeys' performance in this stage. In the experimental stage, all three monkeys succeeded on more than 75% of three-item problems (Avdagic et al. 623-4). Their performance on four-item problems varied considerably between the first and second phases of the experimental stage. For four-item problems in the first phase, the relative frequency of successful trials ranged from 39.2% (MacDuff) to 58.8% (Lashley) for radius trials and from 32.6% (MacDuff) to 40% (Lashley) for luminosity trials. All subjects were significantly more accurate than chance performance of 1.7% (Avdagic et al. 623-4). In the second experimental phase, performance on four-item problems was less accurate but still considerably better than chance. Accuracy ranged from 19.8% to 43.5% on luminosity trials and from 13.5% to 20.8% on radius trials. As in the first phase, Lashley performed most accurately on both tasks; MacDuff performed least accurately on second-phase luminosity problems and Oberon set the low bar for radius problems.

What is important for our purposes is the basic structure of the problem the monkeys faced and the means Avdagic et al. furnished for them to solve that problem. Consider what the monkeys must learn in the training phase in order to perform accurately in the experimental stage. First, they must learn to perform each of the simultaneous chaining tasks individually. Second, they must learn to notice and condition their final behavioral response on the background color of the screen. Third, they must learn to associate each background color with the appropriate task, i.e. to attend to the size of the stimulus items when the background is red and to attend to their brightness when the background is blue.

Success in Avdagic et al.'s task switching setup requires coordinating these attentional and behavioral dispositions to enable the learner to accurately and flexibly switch between task-specific sets of responses when called for by changes in their environment. In training phases 1 and 2, they have the opportunity to learn each task individually in the presence of the cue used to signal task switches in later stages (background color), gradually moving through problems involving successively larger collections of items. And in phase 3, they have a chance to practice sequences of randomly-switching problems.

4.3 Moran Learning

To model the kind of learning accomplished by Avdagic et al.'s monkeys, we need to specify a learning dynamics. Here, we use *Moran learning*, a relatively underexplored reinforcement learning dynamics introduced in Barrett (2006), to model the monkeys' learning. This dynamics treats learning as a *Moran process*, a kind of stochastic process commonly used in biological models of evolution in finite populations (see Moran (1958)). Moran learning can be modeled with balls and urns. Suppose an agent repeatedly faces a choice problem in which it must choose one of k many acts $a_1, a_2, ..., a_k$. The agent chooses by drawing a ball at random and without bias from an urn containing balls labeled with numbers from 1 to k. If it draws a ball with an i label ($1 \leq i \leq k$), then it chooses a_i.

Suppose the agent faces n trials of the choice problem, and let t_j stand for the jth repetition in the sequence. Suppose further that the urn initially contains c many balls. We assume that the outcomes of the learner's choice are bivalent: a given choice results in either *success* or *failure*. If the agent is a Moran learner, then it updates the urn's contents in the following way. If the agent chooses a_i in t_j and succeeds, then it first *removes* r ($r < c$) many balls at random.[4] Then r many balls of type i are added to the urn. If instead the agent's choice of

[4]We assume the urn is well-mixed so that, for each ball removed and each label-type i, the probability

a_i in t_j results in failure, then the contents of the urn remain unchanged. If the agent learns by Moran learning *with punishment*, then successful actions are reinforced in the way just described, but if a chosen action a_i results in failure, the learner first removes p many balls of the unsuccessful type (where $p < c$) and then adds p many balls such that for each type l, the probability that a given newly-added ball is an l-type is equal to the proportion of l balls in the urn after the initial removal of balls.

The variant of Moran learning with punishment used in our model involves a small modification to the description above. If removing balls in either a reinforcement or punishment event reduces the number of balls of some type in the urn to zero, then a new ball of that type is added and a random ball of a different type (which has more than one representative in the urn) is removed. This ensures that no act type goes to extinction, i.e., that the agent always chooses each act with positive (but possibly very small) probability.

A key motivation for using Moran learning in our context concerns its treatment of *forgetting*. A learning dynamics is *forgetful* if more recent experience exerts a greater influence than less recent experience in determining the agent's choice dispositions.[5] There are numerous ways to introduce forgetting into a reinforcement learning dynamics. One way involves building in an adjustable reference point, where forgetting is introduced by allowing the learner to "get used to" a given payoff level, such that the magnitude of the agent's reinforcement on a given action-outcome pair depends on how often outcomes with similar payoffs have occurred in the recent past.[6] In this model, the risk of lock-in on suboptimal dispositions is countered by the exploration encouraged by the learner's tendency to treat a given objective payoff level as less valuable (and possibly eventually as a positive *loss*) as it comes to reliably receive payoffs at or above that level. Other implementations of forgetfulness involve periodically throwing out a certain number of randomly selected balls from the learner's urn.

---

that that ball is an i-type is equal to the proportion of i-type balls in the urn.

[5]For an extensive discussion of the role of forgetting in learning, see Barrett and Zollman (2009).

[6]See Bereby-Meyer and Erev (1998) for discussion of this kind of learning.

Under Moran reinforcement learning, forgetting is implemented by allowing reinforcement events to replace a constant *proportion* of the learner's urn's contents, rather than simply adding a constant number of balls upon success. This allows a Moran learner to adapt its dispositions to new practical contexts more nimbly than in models in which propensities are unbounded and reinforcements matter cumulatively for choice probabilities (as in, e.g., Roth and Erev's (1995) model). What recommends this learning rule for our purposes is just that it exhibits the basic structure of reinforcement learning with forgetting. The particular mechanism by which forgetfulness is implemented is not theoretically significant. That said, there are pragmatic considerations that favor Moran learning. In particular, it is a relatively simple learning rule, having just three free parameters[7] (the reinforcement and punishment levels, along with the capacity of the urn) with straightforward psychological interpretations. And Moran processes are well understood mathematically and familiar across a range of statistical, biological, and social scientific disciplines. An interesting extension of our project would consider how our model would perform under other forgetful dynamics.

Forgetting is particularly useful for reinforcement learners in task switching problems. In these problems, the practical demands confronting an agent change often, despite that agent facing a qualitatively similar environment in each trial. Of course, the risk of premature lock-in at the level of task-specific behavioral dispositions is mitigated by the availability of a task cue the learner can use to choose which such dispositions to activate on a given trial. But there remains a serious risk of such lock-in at the level of attentional dispositions, assuming that those dispositions also evolve by reinforcement learning. In the lab, subjects are typically trained on each task individually before facing mixed trials, and there is no extrinsic reward for attending to the task cue in early single-task training stages. If a learner habitually ignores the task cue in those stages, then without some form of forgetting, the force of past successes may make it very difficult to correct course and eventually learn to attend to that cue.

[7]This assumes the urn's initial contents are fixed.

4.4 The Model

The model of the monkeys' learning involves three *subagents*, each corresponding to a distinct functional role in determining the learner's final behavioral output. The first is the *metasalience controller*. Recall that in the Avdagic et al. experiment, the background color of the screen in a given trial reliably indicates which task the monkey will be rewarded for executing. If the monkey attends to this feature and notices when it changes, it may learn to switch between tasks precisely when necessary to consistently receive rewards. But it doesn't *have to* pay attention to background color. It might ignore the screen's color on any given trial. In the model, the metasalience controller is the subagent responsible for determining whether the learner attends to background color.

Assuming that the metasalience controller has just these two options—*attend to* or *ignore* background color—amounts to supposing the learner has two available ways of categorizing the repeated problems she faces: she can either treat all stimulus displays as instances of the same kind of problem, or she can distinguish between two kinds of problem corresponding to the two possible background colors. Of course, we could envision more complex environments featuring "distractor" cues the learner must learn to ignore in order to consistently succeed in the problem at hand.[8] Such distractions could be straightforwardly implemented in the model by making more actions available to the metasalience controller, where each action corresponds to a different feature of the environment the agent could attend to in distinguishing problems.[9] For the sake of simplicity, they are excluded here.

[8] A concrete example: Perhaps there is a lamp in the corner of the laboratory within the learner's view. Its state (*on* or *off*) is uncorrelated with the task the learner will be rewarded for performing. Nevertheless, the agent might notice and condition its action on the state of the lamp, effectively categorizing the problems they face as "lamp-on problems" and "lamp-off problems." Consistent success in the problem at hand would require the agent to learn to ignore the lamp and instead attend to background color.

[9] In terms of the balls-and-urns interpretation given below, each distractor feature would correspond to a distinct type of ball in the metasalience urn. For each such feature, the *salience controller* (whose behavior is described in detail below) would be equipped with a distinct urn for each possible state of that feature. Upon drawing a ball from the metasalience urn, the metasalience controller would tell the salience controller to draw from the urn corresponding to the observed state of the feature specified by the type of ball drawn. Then, the dynamics would work exactly as described below for the model without distractor cues.

Recall also that in the experiment the stimulus items presented in each trial differ with respect to two features, radius and luminosity. The *salience controller* determines which of these features the learner will condition on in choosing an order in which to select the stimulus items.[10] If the metasalience controller chooses to attend to background color, then the salience controller conditions this choice on whether the background is red or blue; otherwise, background color is ignored in choosing to attend to luminosity or radius.

The *action controller* is responsible for determining the order in which items are selected. If the salience controller chooses luminosity as the salient feature, the action controller distinguishes among objects according to their rank with respect to luminosity; if the salience controller chooses radius, the action controller distinguishes among objects according to their rank with respect to radius length. The action controller can choose the same object multiple times in a given trial; it must learn that repeated selections lead to failure in the problems at hand. In the model, each sub-agent's dispositions—corresponding to noticing or failing to notice the color cue, selecting a feature of the stimuli on which to condition one's actions, and choosing a behavioral response to those stimuli—evolve by Moran reinforcement learning.[11] The total model of choice is illustrated in figure 1.

For a more complete characterization of the model, consider a single trial in Avdagic et al.'s experiment. For concreteness, suppose it is a trial in the third training phase. First, the target task is selected by a random and unbiased choice between the luminosity task and the radius task, the background color of the screen is set accordingly, and the stimulus items are presented to the learner. (In the first two training phases, the task is chosen

[10]The model of attentional learning in the present paper bears much in common with extant work on the evolution of salience in self-assembling games. For more on salience learning, see Herrmann and VanDrunen (2022), Barrett (2023), and Torsell and Barrett (2024). For an introduction to the framework of self-assembling games, see Barrett and Skyrms (2017) and Barrett (2025).

[11]The assumption the reinforcements and punishments occur for all three subagents in lockstep turns out to be inessential to the model's success. See Appendix figure 1 for results for a variant of the model in which the metasalience and salience controllers' acts are only positively reinforced on successful trials (and propensities are unchanged on unsuccessful trials), but the action controller's dispositions evolve by both reinforcement on successful trials and punishment on unsuccessful ones.

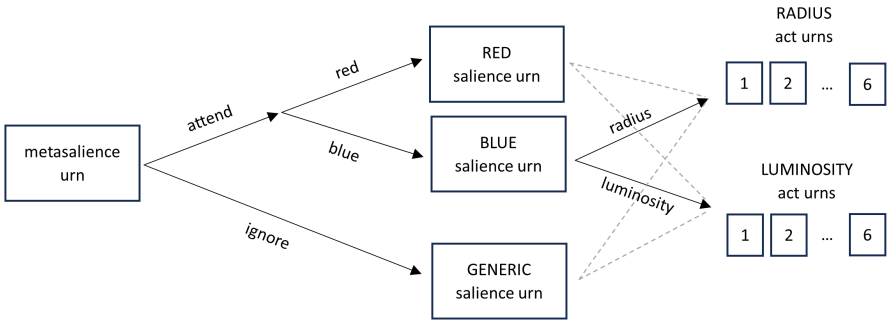

Figure 4.1: The urn model

deterministically.) Then, the number of items is determined. In Avdagic et al.'s experiment, the number of items varies between three and six throughout the third training phase. As a simplifying assumption, our simulations fix the number of items at four for this phase.

Another respect in which our simulated experiment diverges from Avdagic et al.'s setup is in assuming that the stimulus items vary in *both* radius and luminosity throughout all training phases (recall that for Avdagic et al., items vary in only one dimension until the first experimental stage.) In a k-item trial, a random permutation f of $\{1, 2, ..., k\}$ is chosen, where $f(i)$ $(1 \leq i \leq k)$ gives the rank with respect to the distractor dimension of the ith-ranked object with respect to the task-relevant dimension. If, for example, in a three-object radius task we have it that $f(2) = 3$, this is interpreted to mean that the second-smallest stimulus item is also the brightest.

The first step in determining the learner's response to the presented stimuli is executed by the metasalience controller. The metasalience controller draws a ball from a well-mixed urn—the *metasalience urn*—containing balls labeled "attend" and balls labeled "ignore." The metasalience controller's draw determines which of three urns the salience controller will draw from to determine which feature of the stimulus items the learner will condition its final behavioral response on. If an "attend" ball is drawn, then the metasalience controller

observes the background color of the screen and reports the color to the salience controller, who draws from an urn labeled "BLUE" if the background is blue and draws from an urn labeled "RED" if the background is red. Call these the *blue salience urn* and the *red salience urn*. If an "ignore" ball is drawn, then the salience controller draws from an urn labeled "GENERIC". The red, blue, and generic salience urns each contain balls labeled "radius" and balls labeled "luminosity." Note that in the first training phase, the blue salience urn is never drawn from, since the metasalience controller never observes a blue background; similarly, the red salience urn is never active in the second training phase.

The salience controller's draw determines which of two sets of urns the action controller will draw from to determine the order in which to select the stimulus items. One set of urns is labeled "RADIUS" and the other is labeled "LUMINOSITY." Call these the *radius act urns* and *luminosity act urns*. Each set contains six urns, labeled with digits from 1 to 6 subscripted with an r or an l to distinguish urns in the luminosity set from urns in the radius set (so that, e.g., the radius act urns include an urn labeled "2_r" and the luminosity act urns include an urn labeled "2_l"). Each of these act urns contains balls labeled with digits ranging from one to six.

In considering how the action controller determines the final behavioral response, there are two possible cases. The first is the case in which the salience controller draws a ball labeled with the task-relevant dimension—suppose, without loss of generality, that it is luminosity—and so the learner attends to luminosity in choosing an order in which to select the items. In that case, the action controller first draws from the 1_l urn to determine which item it will select first. If the learner draws a 1-labeled ball, then its first selection will be of the least luminous object—the unique task-appropriate choice—and it will then determine its second selection by drawing from the 2_l urn. If that selection is successful (i.e., if the learner draws a 2-ball, and so selects the item ranked second with respect to luminosity), it will make its third selection by drawing from the 3_l urn, and so on. If it correctly ranks all

the presented stimulus items according to luminosity rank by drawing balls whose labels match the corresponding urn labels, then the trial is successful. The agent receives a reward, and *every urn* that was drawn from in that trial is updated by Moran reinforcement with a reinforcement magnitude of r (i.e., first a random selection of r many balls is removed from the urn, then r balls of the type just drawn are added to the urn).

In the second kind of case, the salience controller draws a ball corresponding to the task-irrelevant dimension, and so the learner's behavioral response will be determined by draws from the radius act urns (we maintain the assumption that the task-relevant dimension is luminosity). Then, since the learner attends to the distractor dimension, it will succeed on that trial just if it draws an $f(1)$-ball from the 1_r urn, an $f(2)$ ball from the 2_r urn, and so on (because f relates the rank of a given object with respect to the task-relevant dimension to its rank with respect to the distractor dimension). If the agent does successfully rank all the items, it updates all urns drawn from in that trial by Moran reinforcement just as in the case where the task-relevant dimension is salient.

Regardless of whether the agent attends to the task-relevant or distractor dimension, if it selects an object out of order, punishment proceeds as follows: the final act urn drawn from (i.e., the urn drawn from in determining the incorrect selection), the salience urn drawn from, and the metasalience urn are updated by Moran punishment (i.e., p many balls of the type just drawn are removed and replaced according to the rule described above). Notice that the contents of the act urns drawn from prior to the incorrect selection remain unchanged.

Recall that the number of stimulus items presented (k) varies in some phases of Avdagic et al.'s experiments. In our model, this is accommodated by allowing the action controller to first observe the number of stimulus items on the screen in a given trial and then remove, from all urns labeled with a number no larger than k, all balls labeled with numbers larger than k. Suppose, for example, that the learner faces a trial in which $k = 3$. Since only three stimulus items are presented, the learner will consult only the first three urns in the set of act

urns corresponding to the salience controller's draw. All balls labeled with numbers greater than 3 are removed from these urns. The idea is that the learner's disposition to choose, e.g., the fifth-most luminous item in a collection of stimuli will not be activated in a trial in which it sees just three items; only the learner's propensities to select the first, second, and third brightest stimuli (represented by the number of 1-, 2-, and 3-labeled balls, respectively, in the relevant urn) should count in determining its behavior.

Of course, the model does not capture all potentially relevant features of the monkeys' learning and cognition or all the details of Avdagic et al.'s experimental protocol. It is particularly significant that the radii and luminosities of the stimuli are represented only in terms of their ordinal rank—there is no way to compare the differences between pairs of stimuli with respect to these features. Recall that the experimenters facilitated the monkeys' learning in the first experimental phase by reducing the degree to which the stimulus items varied in the distractor dimension relative to the task-relevant dimension. Since the model represents stimulus items only in terms of their ordinal ranks, it cannot capture this strategy for inducing an attentional bias in favor of the task-relevant dimension. Nor does the model enable us to represent how reduced differences in the distractor dimension relative to the task-relevant dimension would influence the monkeys' learning, as the differences between the monkeys' performance in the first and second phases of the experimental stage clearly indicate it did.

Furthermore, in all of Avdagic et al.'s training phases, it is possible for the monkeys to learn to perform accurately by conditioning their behavior on *which feature varied across stimulus objects in a given trial*, rather than background color, since stimuli varied only with respect to the task-relevant dimension throughout the training stage. For simplicity, no parameter for the salience of this feature of the subjects' environment is included. As a result, the model cannot capture how limiting variation to the task-relevant dimension might assist the monkeys' learning, and so the model assumes that objects vary with respect to both

dimensions in each trial of the training stage.

These simplifying assumptions—that stimulus items are fully characterized by their ordinal rank with respect to the two dimensions and that the items presented always vary in both dimensions—have the consequence that the model cannot capture the differences between the protocols of the third training phase and those of the two experimental phases. In the model, learning to successfully navigate the final training phase *just is* learning to flexibly switch between the two SimChain tasks in the constant presence of variation with respect to both luminosity and radius, with only background color indicating which task will be rewarded in a given trial; in Avdagic et al.'s experiment, the monkeys did not face "complete" task switching problems of this kind until the experimental stage. For this reason, simulations were run only on the three simulated training phases.

4.5 Results and Discussion

Simulations were run in which the learner faced the three training phases in succession. As in Avdagic et al.'s experiment, the learner progressed by successively meeting the 80% performance criterion on problems with increasingly large collections of stimulus items. A simulated run was counted as a success if the learner progressed through both single-task training phases and reached the 80% criterion in the third phase.

Initially, the salience urns and the metasalience urn each contained 100 balls (50 of each possible type). Similarly, each of the action controller's urns contained 50 balls of each possible type (for a total of 300 balls in each urn). Recall the structure of the first two training phases: subjects progress through blocks of problems with an increasing number of stimuli as they successively meet the performance criterion. Because the action controller ignores balls and urns labeled with numbers greater than the number of stimulus items

presented, we can think of each of the first two training phases as beginning with the action controller working with just three urns in each set (i.e., the luminosity act urns and the radius act urns), each of these containing 150 balls (initially: fifty 1-balls, fifty 2-balls, and fifty 3-balls). Each time the learner meets the performance criterion for the k-item task, a new urn is added to both the luminosity and radius sets (containing fifty balls for each label from 1 to $k+1$), and fifty k+1-labeled balls are added to each of the k act urns that were available in the preceding trial, until the criterion is met for the maximum number of items.

Simulations were executed in batches of 1000 runs for all reinforcement and punishment values $(r, p) \in \{1, 2, 3\} \times \{0, 1, 2, 3\}$. For each batch, r and p were held constant across and within runs. On all parameter values tested except $(1, 0)$, all runs ended in success. The results reported in this section are for 1000 runs of the entire training sequence with $r = 2$ and $p = 1$. There is nothing special about these values. While the speed of learning varied across parameter settings, with higher values for both r and p corresponding to faster learning, the qualitative patterns reported below held across all parameter settings tested. For results on alternative parameter settings, see Appendix figure 2.

training phase	items/problem	mean no. sessions to 80% criterion
1	3	12.73
	5	9.96
2	3	62.72
	5	10.07
3	4	7.6

Table 4.1: mean time-to-criterion for each simulation phase

Figure 2 reports the mean number of 50-trial sessions it took for the agent to reach the performance criterion for each phase of the training stage. Avdagic et al. do not report analogous results for the training phases of their experiments, so no direct comparison between the performance of the model and the performance of the monkeys is possible. However,

it is worth noting that the values for average sessions-to-criterion for our simulations never exceed 63—and in fact, with the exception of the first segment of phase 2 (which is discussed in greater detail below), all values are less than 13. Given that the experimental phases consisted of 36 and 48 fifty-trial sessions each, this suggests that, at least on our chosen parameter settings, the modeled agent is able to progress through the training stage within a reasonable timeframe for an experiment of this kind.

Notice that, although the number of stimulus items increases in each successive segment of the first two phases, the number of sessions-to-criterion did not uniformly increase as more items were added. This is because the structure of the model allows for significant transfer of learning from earlier segments to later ones. Once the learner has evolved up dispositions allowing it to achieve $\geq 80\%$ accuracy on, say, three-item radius problems in phase one, it has already achieved much of the learning necessary to reliably complete four-item problems. It is already disposed to choose a "radius" ball from the salience urn with high probability, to choose a "1"-ball from the 1_r urn with high probability, to choose a "2"-ball from the 2_r urn with high probability, and to choose a "3"-ball from the 3_r urn with high probability.

Moving into four-item problems, the learner does face some new challenges: the fifty "4" balls added to the three act urns used in the previous segment interfere with its evolved dispositions, and the contents of the newly-activated 4_r urn are uniform across ball types (meaning that initially the learner will make a correct fourth-item selection with chance 0.25). But the learning the agent has already accomplished is a powerful aid to meeting these challenges. In particular, the favorable dispositions built up in the three previously-active act urns, even when diluted by the new "4" balls, give the learner many more opportunities to attempt, and learn from, drawing from the 4_r urn than it initially would have were it to start with uniform proportions of types in all the relevant act urns. Were the agent to enter the four-item problems with its untrained uniform propensities, it would have the opportunity to make a fourth-item selection (i.e., draw from its fourth act urn) only on those trials in

which its random guessing produced three correct choices in succession.

The agent's learning in the first two training phases thus exhibits a kind of *bootstrapping*, in which the learning accomplished in earlier, simpler problems facilitates faster learning in later, more difficult problems. The structure of these phases—that the number of items subjects are required to rank gradually increases from an initially small number—is important to the agent's success in working up to the task switching problems of the third phase insofar as it allows the agent to take advantage of this learning transfer.

A striking feature of the results in figure 2 is the large difference in time-to-criterion between the first segment of the first training phase and the first segment of the second training phase. The explanation of this feature has to do with the metasalience controller's dispositions. Figure 3 plots, for each run, the final probability with which the learner attended to background color for each of the three training phases, given by the proportion of "attend" balls in the metasalience urn at the end of each phase. Results are reported at a precision of 0.05. Notice that the distribution for phase 1 is bimodal, with peaks close to zero and close to one. On about half of runs, the learner had either evolved to reliably attend to background color or to reliably ignore it by the time it reached the performance criterion for six-item radius problems. For the other half, the learner exited the first phase with an intermediate probability of attending to background color.

How reliably the learner had evolved to attend to background color by the end of the first phase of a run strongly influenced the difficulty of adjusting to the new practical demands of phase 2 on that run. When the learner evolves to reliably attend to background color, working up to the performance criterion for the three-item problems of phase 2 is no more challenging than reaching the criterion in the first segment of phase 1. But when the learner evolves to ignore background color, considerable new difficulties arise. In this case, the learner draws from the same set of salience and act urns she had trained up in the first stage, saddling her with a strong bias in favor of attending to the task-irrelevant dimension.

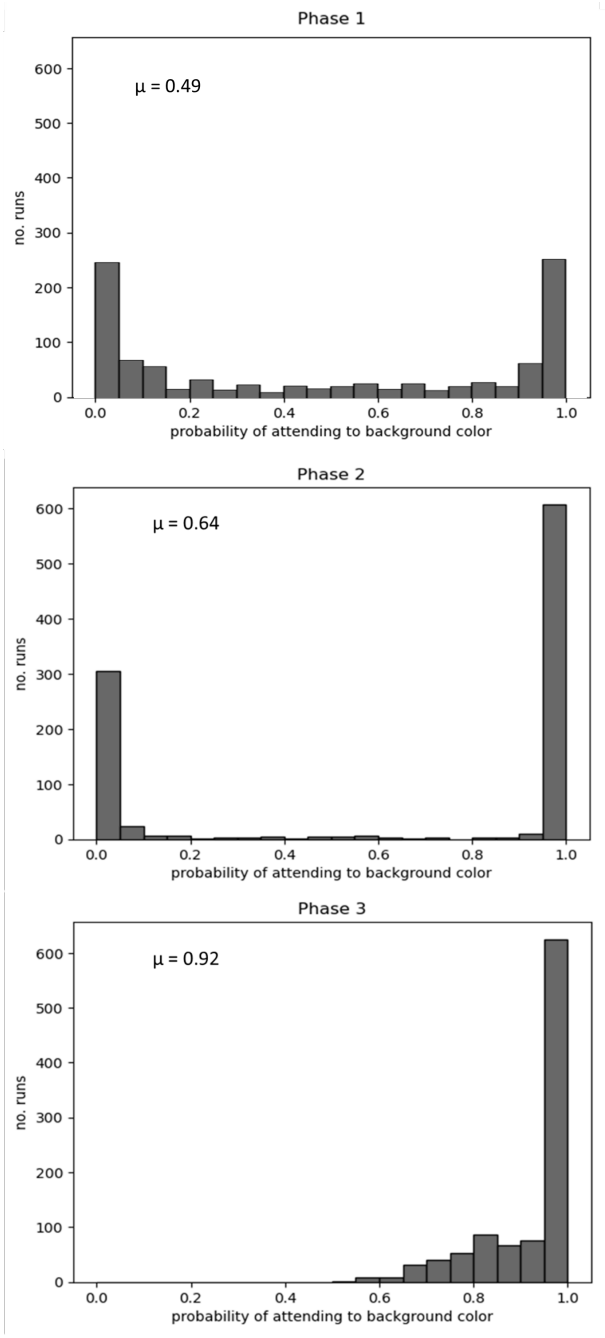

Figure 4.2: Final metasalience choice probabilities, phases 1-3

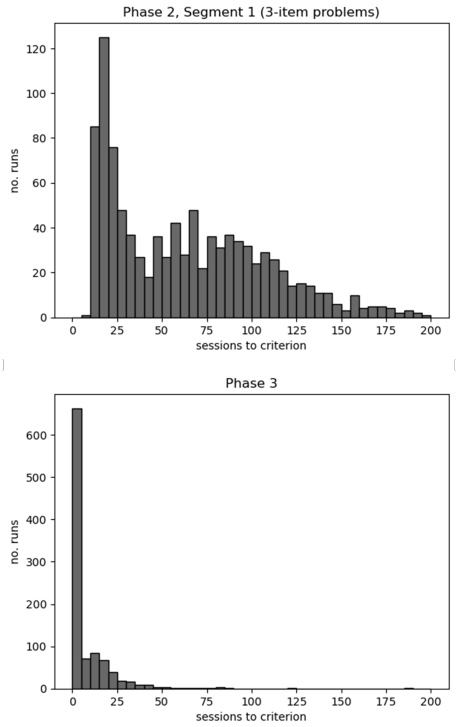

Figure 4.3: Sessions-to-criterion for phase 3 and the first segment of phase 2

She must retool to meet the demands of the new environment by either gradually unlearning the bias of the generic salience urn or else learn to attend to the background color and develop task-appropriate attentional dispositions via the blue salience urn. This is what explains the large mean sessions-to-criterion value for the first segment of phase 2.

There are a few more things to note about the metasalience probabilities recorded in figure 3. The distribution of final metasalience probabilities after phase 2 indicates a shift in favor of attending to background color, with nearly twice as many runs resulting in a probability close to one than resulting in a probability close to zero. Compared with the data for phase 1, the mean probability of attending to background color has increased by nearly 0.15. Additionally, the number of runs ending with an intermediate probability of attending to background color has decreased.

As in the transition from phase 1 to phase 2, the probability with which the learner attended to background color influenced the difficulty of learning to adjust to the new challenges of

phase 3. When the learner exited phase 2 disposed to attend to background color with high probability, it entered phase 3 already disposed to condition its behavior on the feature of its environment that would reliably signal which task it would be rewarded for performing in phase 3. On runs in which the agent had learned to attend to background color with high probability in *both* the first and second phase, it entered phase 3 with a strong metasalience bias in favor of attending to background color, a strong salience bias in favor of attending to the task-relevant dimension conditional on the background color, and act-level dispositions accurate enough to successfully complete the training phases for each individual SimChain task—precisely the dispositions necessary for reliable success in phase 3. In the time-to-criterion distribution shown in the bottom panel of figure 4, these are the runs accounting for the large spike in the leftmost bin of the histogram. By contrast, when the learner exited phase 2 disposed to ignore background color with high probability, it had to *learn* to notice shifts in background color and to associate each color with the appropriate set of salience- and act-level dispositions in the third phase. The maximum time-to-criterion for phase 3 was 185 sessions, though in the overwhelming majority of runs, the phase was completed in fewer than 50 sessions.

Overall, these results indicate that the model reliably learns to cope with the inductive challenges distinctive of task switching problems with a high degree of success. It is important to emphasize again the simplicity of the model's basic structure: three simple subagents responsible for selecting what the learner pays attention to and how it behaves evolve by reinforcing successful dispositions and punishing unsuccessful ones. In a simulated environment in which the learner first has an opportunity to learn two distinct ranking tasks in isolation and then confronts a series of trials that randomly alternate between the two tasks, the operation of this simple dynamics gradually assembles a system that can deftly navigate the task switching problems with a high degree of success. The model thus illustrates one way in which a learner might accomplish what Avdagic et al.'s subjects achieved in the lab. It is significant that the model's success arises from a simple kind of reward-based association

learning operating on the agent's attentional and behavioral dispositions.

4.6 Conclusion

Natural learners face a problem of projectibility inasmuch as they must learn which regularities in their environment are useful guides for action and prediction in order to survive. As in Chapter 3, we considered a concrete laboratory setting in which subjects face a version of this problem, using a model to show how that problem might be solved by means of a simple form of reinforcement learning. In the model, the dynamics operates simultaneously on two kinds of attentional dispositions and first-order behavioral dispositions. As these dispositions coevolve, the agents learn to perform two distinct tasks and to selectively activate the dispositions suited to each task depending on the state of a task cue. On simulation, the learner reliably comes to perform both tasks with a high degree of accuracy and to attend to the task cue to distinguish between them.

The model thus shows how the problem of projectibility that arises in task-switching problems might be solved by means of a particularly simple form of learning. An important virtue of this model is that it describes a learning process in which the mechanism responsible for learning which regularities to project—salience learning by means of reinforcement—operates *simultaneously* with the mechanisms responsible for learning the first-order tasks the agent must learn to discriminate. That is, the agent learns which regularities in its environment to project for the purpose of distinguishing the tasks even as it learns the tasks themselves. What's more, all the relevant dispositions evolve by the *same* simple reinforcement mechanism.

The resulting story is one on which a highly structured learning system emerges endogenously from what is initially a largely unstructured collection of attentional and behavioral

dispositions. The model might thus be interpreted as capturing the *self-assembly* of a learning method adapted to the demands of the task-switching problem at hand. In particular, it captures a process that reliably assembles a learning method that projects the practically relevant regularities in a particular kind of task-switching problem.

Of course, not all the relevant structure is learned endogenously. Some is baked in. We assume, for example, that our learner starts out "knowing that" either brightness or size is what matters in the problem at hand, that she is not distracted by potential cues other than background color in individuating tasks, and that she represents stimulus items according to their order along the two relevant dimensions. The self-assembly involved in the model is not self-assembly *ex nihilo*. An interesting question for further work concerns how the structural features of the agent stipulated at the start might themselves be learned or evolved. Moreover, we studied just one form of reinforcement learning as the mechanism of self-assembly. Another useful step in the direction of greater generality would be to investigate whether other forms of reinforcement learning would be similarly successful.

Chapter 5

Conclusion

Problems of projectibility are ubiquitous in nature. Successful learning requires selective sensitivity to certain regularities in experience—the ones that are in fact reliable guides for action and prediction—and insensitivity to the rest.

Above, we considered how simple agents might learn by trial and error to attend to projectible regularities in problems involving signaling, perceptual discrimination, and task switching. These are kinds of problems that confront a wide variety of natural learners at varying scales of biological and cognitive complexity.

Throughout, the notion of self-assembly has played a central role. In particular, the models we discussed capture processes in which an effective method of learning gradually emerges out of the initially random behavior of a collection of simple functional components of a learner. The mechanisms of self-assembly we have considered are simple kinds of reinforcement learning operating on those components' probabilistic dispositions. As the components themselves learn, their interactions crystallize into a composite learning system that can reliably solve a particular kind of problem. This is the self-assembly of learning. Following our central theme, we have been particularly attentive to the inductive assumptions of pro-

jectibility reflected in the evolved structure of the learner.

While we have emphasized the endogenous emergence of structure, each model necessarily begins with *some* structure imposed by the modeler. In constructing any model of learning, one must make choices in specifying the set of available actions, the rules by which an action is selected at a given time, and the dynamics governing the evolution of choice dispositions. These choices reflect substantive assumptions about the cognitive and behavioral dispositions and capacities the learner brings to a problem. At the end of the last section, we highlighted some of the assumptions that went into the task switching model. We could similarly catalog the assumptions built into the other models' initial setups.

There is of course no hope of modeling self-assembly from nothing. Instead, our goal has been to construct models that incorporate only capacities and dispositions that are widely available to learners in nature and which can implemented with meager cognitive resources. This is what motivates the use of simple forms of reward-based learning, like Roth and Erev's rule. And it is itself motivated by the kind of naturalism illustrated by Hume and Suppes, which sees projectibility as a concern for all natural learners and thus seeks explanations for successful projection that apply to organisms other than adult humans making explicit inferences.

Moreover, these models suggest that the basic mechanisms of self-assembly they invoke could be used to explain the emergence of features built in by the modeler upfront. That is, they suggest that one could take any particular exogenously fixed disposition or capacity in one of our models and show how it might evolve from a less structured starting point under the same kinds of trial-and-error processes driving the original model. As the discrimination learning model illustrates, even the learning dynamics itself can self-assemble. The idea is that, while self-assembly *ex nihilo* is an impossibility, we can push back the starting point of our evolutionary story by leaving more and more features of the target system to emerge under the dynamics.

We introduced issues of projectibility by revisiting their best-known appearance in the philosophical literature, in Nelson Goodman's new riddle of induction. But the subsequent discussion veered far from Goodman's original concerns. Goodman was interested in the logic of induction. In working out what it takes for a universal statement to be genuinely confirmed by its instances, he found he'd need an account of how to distinguish the predicates that could legitimately figure in such a statement from those that could not. That is, he needed a criterion for identifying the predicates that standard inductive practice licenses us to project in inductive inference. We instead have been concerned with how inductive assumptions concerning projectibility might be *learned*. In particular, we have been concerned with how such assumptions might be learned by primitive reinforcement learning methods.

To return to another figure from our introductory discussion, we have taken our cue from Hume's skeptical solution to the old problem of induction. The models discussed above are not meant to show that the learners they represent are justified in projecting the regularities they come to attend to. Rather, they show that simple associative learning suffices to guide them to project regularities that are in fact reliable guides to successful action in the problems they repeatedly face. Of course, there is no *guarantee* that the patterns they look for will continue indefinitely to provide helpful practical guidance. The modeled learners are nevertheless disposed to persist in projecting them because doing so has led to successful action in the past and their psychological nature disposes them to repeat operations that have been rewarded.

Bibliography

J. McKenzie Alexander, Brian Skyrms, and S. L. Zabell. Inventing new signals. *Dynamic Games and Applications*, 2(1):129–145, 2012.

R. Argiento, R. Pemantle, Brian Skyrms, and S. Volkov. Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications*, 119(2): 373–390, 2009.

E. Avdagic, G. Jensen, D. Altschul, et al. Rapid cognitive flexibility of rhesus macaques performing psychophysical task switching. *Animal Cognition*, 17:619–631, 2014. doi: 10.1007/s10071-013-0693-0.

Jeffrey A. Barrett. Numerical simulations of the Lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar. Technical Report 54, Institute for Mathematical Behavioral Sciences, 2006. Paper 54.

Jeffrey A. Barrett. Self-assembling games and the evolution of salience. *The British Journal for the Philosophy of Science*, 2023. doi: 10.1086/714789. URL `https://www.journals.uchicago.edu/doi/10.1086/714789`.

Jeffrey A. Barrett. Humean learning (how to learn). *Philosophical Studies*, 181(1):281–297, 2024.

Jeffrey A. Barrett. *Self-Assembling Games: Language, Learning, and Inquiry*. Oxford University Press, 2025. Forthcoming.

Jeffrey A Barrett and Travis LaCroix. Epistemology and the structure of language: Ja barrett, t. lacroix. *Erkenntnis*, 87(2):953–967, 2022.

Jeffrey A. Barrett and Brian Skyrms. Self-assembling games. *The British Journal for the Philosophy of Science*, 68(2):329–353, 2017.

Jeffrey A. Barrett and Kevin J. S. Zollman. The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence*, 21(4): 293–309, 2009.

Alan W. Beggs. On the convergence of reinforcement learning. *Journal of Economic Theory*, 122(1):1–36, 2005.

Yoella Bereby-Meyer and Ido Erev. On learning to become a successful loser: A comparison of alternative abstractions of learning processes in the loss domain. *Journal of Mathematical Psychology*, 42(2–3):266–286, 1998.

Robert R Bush and Frederick Mosteller. A mathematical model for simple learning. *Psychological review*, 58(5):313, 1951.

Robert R. Bush and Frederick Mosteller. *Stochastic Models for Learning*. John Wiley & Sons, New York, 1955.

Ido Erev and Greg Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112(4):912, 2005.

Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88:848–881, 1998.

Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

Nelson Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1983.

Harry F. Harlow. The formation of learning sets. *Psychological Review*, 56(1):51–65, 1949.

Daniel Herrmann and Jack VanDrunen. Sifting the signal from the noise. *The British Journal for the Philosophy of Science*, 2022.

Richard J. Herrnstein. On the law of effect. *Journal of the Experimental Analysis of Behavior*, 13(2):243–266, 1970.

Josef Hofbauer, Karl Sigmund, et al. *Evolutionary games and population dynamics*, volume 1. Cambridge university press Cambridge, 1998.

David Hume. *An Enquiry Concerning Human Understanding*. Oxford University Press, 2007.

Gerhard Jäger. The evolution of convex categories. *Linguistics and philosophy*, 30(5):551–564, 2007.

Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. Control and interference in task switching—a review. *Psychological Bulletin*, 136(5):849, 2010.

Karl S. Lashley. *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain*. University of Chicago Press, 1929.

Donald H. Lawrence. Acquired distinctiveness of cues: I. transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, 39(6): 770, 1949.

David Lewis. *Convention: A Philosophical Study*. John Wiley & Sons, 1969.

Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, 2003.

P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *American Economic Review*, 85:1313–1326, 1995.

Cailin O'Connor. Evolving perceptual categories. *Philosophy of Science*, 81(5):110–121, 2015.

Cailin O'Connor. The evolution of vagueness. *Erkenntnis*, 79(Suppl 4):707–727, 2014.

Cailin O'Connor. Games and kinds. *The British Journal for the Philosophy of Science*, 70 (3):719–745, 2019.

Robert A. Rescorla and Allan R. Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black and William F. Prokasy, editors, *Classical Conditioning II: Current Research and Theory*, pages 64–99. Appleton-Century-Crofts, New York, 1972.

Robert W. Rosenthal. Rules of thumb in games. *Journal of Economic Behavior & Organization*, 22(1):1–13, 1993.

Alvin E. Roth and Ido Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8(1): 164–212, 1995.

Brian Skyrms. *Signals: Evolution, Learning, and Information*. Oxford University Press, 2010.

Brian Skyrms. *From Zeno to Arbitrage: Essays on Quantity, Coherence, and Induction*. Oxford University Press, 2012.

Dale O. Stahl. Boundedly rational rule learning in a guessing game. *Games and Economic Behavior*, 16(2):303–330, 1996.

Patrick Suppes. Learning and projectibility. In Douglas Stalker, editor, *Grue!* Open Court Publishing, 1994.

N. S. Sutherland and N. J. Mackintosh. *Mechanisms of Animal Discrimination Learning*. Academic Press, 1971.

Edward L. Thorndike. *Animal Intelligence: Experimental Studies*. Macmillan Press, 1911. doi: 10.5962/bhl.title.55072.

Christian Torsell. Task switching and natural projectibility. *Synthese*, 207(1):51, 2026.

Christian Torsell and Jeffrey A. Barrett. Learning how to learn by self-tuning reinforcement. *Synthese*, 203(6):209, 2024.

Thomas R. Zentall and Donald A. Riley. Selective attention in animal discrimination learning. *The Journal of General Psychology*, 127(1):45–66, 2000. doi: 10.1080/00221300009598570.

Thomas R Zentall, Edward A Wasserman, Olga F Lazareva, Roger KR Thompson, and Mary Jo Rattermann. Concept learning in animals. *Comparative Cognition & Behavior Reviews*, 3, 2008.